SANDIA REPORT

SAND2015-2461
Unlimited Release
Printed April 2015

Modeling an Application's Theoretical Minimum and Average Transactional Response Times

Mary Rose Paiz

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2015-2461
Unlimited Release
Printed April 2015

Modeling an Application's Theoretical Minimum and Average Transactional Response Times

Mary Rose Paiz

Application Services and Analytics

Sandia National Laboratories

MS 1465

Albuquerque, NM 87123

Abstract

The theoretical minimum transactional response time of an application serves as a basis for the expected response time. The lower threshold for the minimum response time represents the minimum amount of time that the application should take to complete a transaction. Knowing the lower threshold is beneficial in detecting anomalies that are results of unsuccessful transactions. On the converse, when an application's response time falls above an upper threshold, there is likely an anomaly in the application that is causing unusual performance issues in the transaction. This report explains how the non-stationary Generalized Extreme Value distribution is used to estimate the lower threshold of an application's daily minimum transactional response time. It also explains how the seasonal Autoregressive Integrated Moving Average time series model is used to estimate the upper threshold for an application's average transactional response time.

Acknowledgment

I am thankful to my manager, Greg N. Conrad, and members of the Application Services and Analytics Department, for their constant guidance and support throughout my research and analysis.

Contents

List of Figures

List of Tables

Nomenclature

pdf Probability Distribution Function

cdf Cumulative Distribution Function

ACF Autocorrelation Function

KPSS Kwiatkowski-Phillips-Schmidt-Shin Test

EVT Extreme Value Theory

EVTS Extreme Value Theorem for Stationary Time Series

AR Autoregressive

MA Moving Average

ARMA Autoregressive Moving Average

ARIMA Autoregressive Integrated Moving Average

Chapter 1

Introduction

Motivation

Sandia National Laboratory's Application Services and Analytics department's middleware services directly support more than 350 enterprise applications with more than 9,500 distinct users during normal business hours. Since their enterprise applications provide support to the entire laboratory, it is important that they constantly monitor and improve Sandia's Enterprise Information System. This is achieved, not only by maintaining a superior level of reliability, utility, and expediency in their work, but by researching and implementing analytic capabilities that can improve our understanding of an applications transactional response time.

The theoretical minimum transactional response time of an application serves as a basis for the application's expected transactional response time. The first goal of this analysis is to estimate the theoretical daily minimum response time of an application. Estimating the daily minimum response time of an application will result in an estimated lower limit for the application's minimum response time. This lower limit can serve as the lower threshold for an alert system that detects anomalies in the applications. The lower threshold represents the minimum amount of time that the application should take to complete any transaction on that day. The moment a response time falls below this lower threshold, an alert can be sent out notifying that the application is experiencing a performance issue that is resulting in unsuccessful transactions.

When an application's response time is greater than a certain threshold, there is likely an anomaly in the application that is causing unusual performance issues. The theoretical average response time can be used to calculate the value of this threshold. Therefore, the second goal of this analysis is to estimate both the average daily and average hourly response times for an application. The upper limits for these estimates will serve as the upper thresholds that will be used to detect significantly large response times. The upper threshold for the average daily response time will be used to perform end of day problem management; if the observed daily average response time is greater than the upper threshold, then an alert will be sent out notifying that the application had a performance issue throughout the day. The

upper threshold for the average hourly response time will be used to check for anomalies every hour.

Middleware Applications

Oracle eBusiness Suite Application

Oracle is an enterprise resource planning tool that is comprised of financial, supply chain, and project accounting applications that are used daily by the administration at Sandia National Laboratories. For this analysis, we will analyze the response times for a standardized transaction from the Oracle eBusiness Suite application. The methodology used to estimate the theoretical minimum and average response times and the results for the Oracle eBusiness Suite are explained in detail in Chapter 5. Figure 1.1 represents the anatomy of a standardized transaction for the Oracle eBusiness Suite application. It shows how the response time for a transaction is measured.

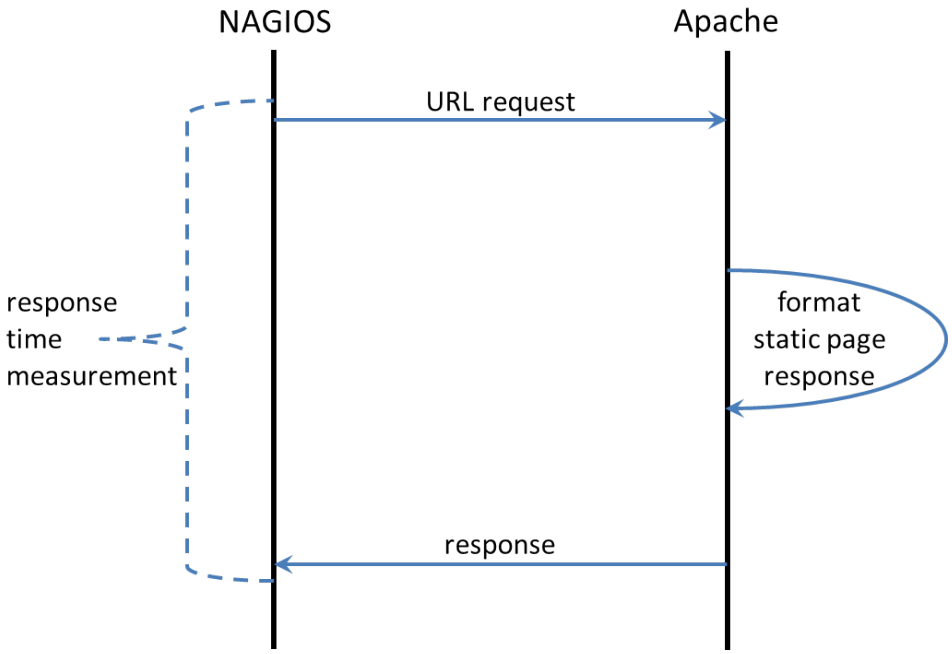


Figure 1.1: Anatomy of a standardized transaction for Oracle eBusiness Suite.

Weblogic 12c Server Application

Weblogic 12c is a Java Enterprise Edition application server. For this analysis, we will analyze the response times for a standardized transaction from the Weblogic 12c Server application. Since the methodology used to perform this analysis is identical to the methodology

used to perform the analysis on the Oracle eBusiness Suite application, only the results are discussed in Chapter 6. Figure 1.2 represents the anatomy of a standardized transaction for the Weblogic c12 Server application. It shows how the response time for ae transaction is measured.

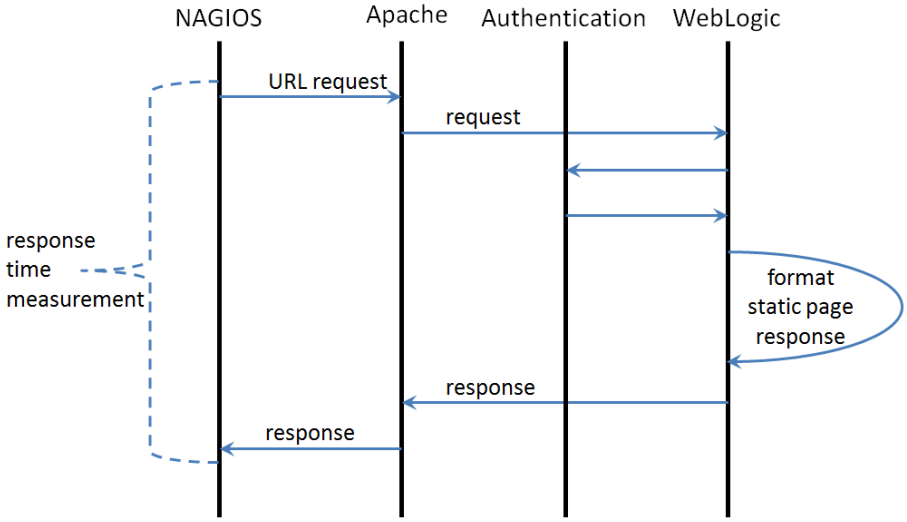


Figure 1.2: Anatomy of a standardized transaction for Weblogic c12 server application.

Although we are only analyzing the transactional response times for the Oracle eBusinnes suite and Weblogic c12 sever applications, the statistical methodology that is used in this analysis can be applied to any application with a transactional system.

Method

The transactional response times that are collected for each middleware application are measured at successive and equally spaced points in time. In statistics, a set of data points that are indexed by time is called a time series. Statistical methods that are used to analyze time series data are known as time series analysis. An introduction to time series analysis can be found in Chapter 2 and the methods that are used in this analysis can be found in Chapters 3 and 4. The programming language that is used to perform this analysis is **R**. **R** is a free programming language and software environment for statistical computing and graphics [11].

The probability distribution function (pdf) that is used to estimate the distribution of the daily minimum transactional response time for an application is the Generalized Extreme Value Distribution (GEV). The theory behind the Generalized Extreme Value Distribution and the parameter selection process is discussed in Chapter 3. The time series model that is used to estimate the daily and hourly transactional response times for an application is

the Autoregressive Integrated Moving Average model ($ARIMA$). The theory behind this model and the selection of the model's orders are discussed in Chapter 4. The results of applying these time series methods to the Oracle eBusiness Suite and Weblogic c12 Server applications are discussed in Chapters 5 and 6, respectively.

Chapter 2

Time Series

Introduction

A data set that is a set of observations that are collected sequentially in time is known as a time series [10]. A time series that has only one observation per time index is known as a univariate time series. For example, let $y_{1:365} = \{y_1, y_2, ..., y_{365}\}$ be the set of data points that represent the daily average transactional response times for an application. The data point y_t denotes the observed average response time for day t. Since the average response times are collected at equally spaced time points, the set $y_{1:365}$ is an equally spaced univariate time series. The methods used in this analysis can only be applied to an equally spaced univariate time series. Figure 2.1 is a plot of the time series $y_{1:365}$.

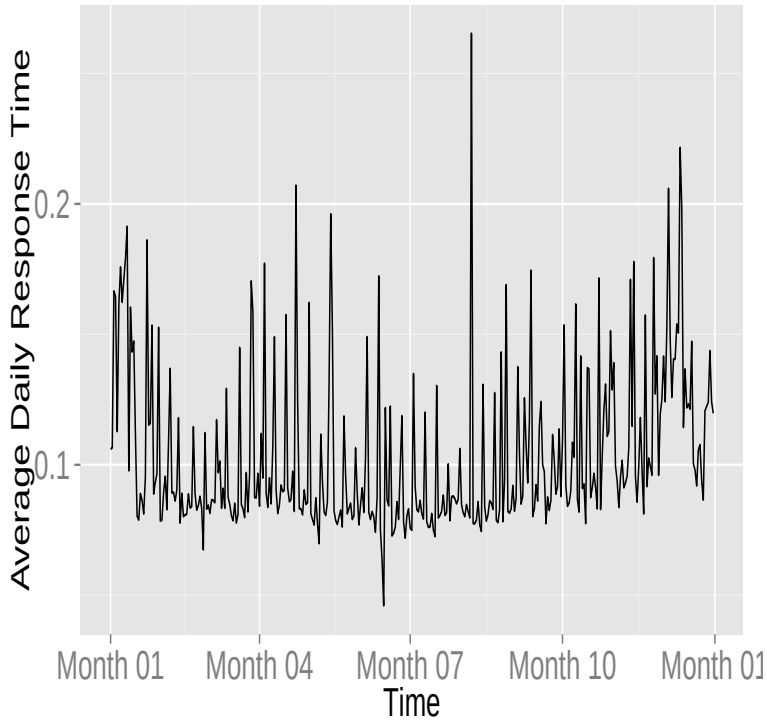


Figure 2.1: An application's average daily response time for $t = 1, 2, ..., 365$.

In mathematics, a random variable is a variable whose value can take on any value in it's sample space. Each value in a random variable's sample space is associated with a certain probability that is based on the random variable's pdf. For example, let $\boldsymbol{Y}$ be the random variable that represents the average daily transactional response time for an application and let Ω denote its sample space. Since response times are always greater than 0, the sample space for $\boldsymbol{Y}$ is $\Omega = \{(0, +\infty)\}$. If $\boldsymbol{Y}$ has a standard uniform distribution ($\boldsymbol{Y} \sim uniform(a = 0, b = 1)$), then all values in Ω have the equal probability of occurrence. Whereas, if $\boldsymbol{Y}$ is normally distributed with a mean equal to 2 and a variance equal to 1 ($\boldsymbol{Y} \sim normal(\mu = 2, \sigma^2 = 1)$), the values in Ω that are close to 2 have a higher probability of occurring [2].

Let $\boldsymbol{X}$ be a random variable that has a pdf denoted by $f_{\boldsymbol{X}}(x)$. The expected value of $\boldsymbol{X}$, is the average value of $\boldsymbol{X}$, weighted according to $f_{\boldsymbol{X}}(x)$. It is denoted by $E[\boldsymbol{X}]$ and defined by

$$E[\boldsymbol{X}] = \int_{\Omega} x f_{\boldsymbol{X}}(x) dx. \tag{2.1}$$

The variance of $\boldsymbol{X}$ is the second central moment of $\boldsymbol{X}$. It is denoted by $Var[\boldsymbol{X}]$ and is defined by

$$Var[\boldsymbol{X}] = E[(\boldsymbol{X} - E[\boldsymbol{X}])^2]. \tag{2.2}$$

The variance of a random variable can be interpreted as the measure of spread of its distribution around its mean [2].

A time series process is a set of random variables that are indexed by time and is denoted by

$$\boldsymbol{Y}_{1:T} = \{\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_T\}.$$

A realization of a time series process is a set of observed values from a time series process. Since we have defined $\boldsymbol{Y}$ as the random variable that represents the average daily transactional response time for an application, the time series plotted in Figure 2.1, $y_{1:365} = \{y_1, y_2, ..., y_{365}\}$, is a realization of the time series process $\boldsymbol{Y}_{1:365}$. For this analysis, a realization of time series processes will be referred to as a time series [10].

A random sample is a set of random variables that have a common probability distribution function, are independent and is denoted by $\boldsymbol{X}_{1:n} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, .., \boldsymbol{X}_n\}$. Let $x_{1:n} = \{x_1, x_2, ..., x_n\}$ be a realization, i.e. set of observed values, from the random sample $\boldsymbol{X}_{1:n}$. An important assumption when fitting observed data to a statistical model is that the data is a realization of independent random variables [2]. This assumption is always met when

analyzing $x_{1:n}$ because the random variables in $\boldsymbol{X}_{1:n}$ are independent. However, this assumption is never met when analyzing a time series because the random variables in a time series process are indexed by time, making them dependent of each other [10].

For this analysis, random variables, random samples and time series process are denoted by a bold capital letter. Observed values are denoted by lower case letters. The sets of letters $(x, \boldsymbol{X})$ and $(y, \boldsymbol{Y})$ are used to differentiate between a random sample and a time series process. For example, $\boldsymbol{X}$ represents a random variable from a random sample and x represents an observation of a random sample. Whereas, $\boldsymbol{Y}$ represents a random variable from a time series process and y represents an observed value from a time series.

Stationarity

An important assumption made when fitting a time series to a statistical model is stationarity. A time series is stationary if it is a realization of a stationary time series process. A time series process is stationary if there is a constant mean and variance throughout the whole process and it's behavior does not depend on when one starts to observe the process [10].

Stationarity in a Time Series Process

Two degrees of stationarity that are widely used in time series analysis are strong stationarity and second order stationarity. It is very difficult to prove that a time series is a realization of a strong stationary time series process. Therefore, we will assume that a time series is stationary if we can prove that it is a realization of a time series process that is second order stationary. A time series process, $\boldsymbol{Y}_{1:T} = \{\boldsymbol{Y}_1, ...\boldsymbol{Y}_T\}$, is said to be second order stationary if for any sequence of times, $t_1, .., t_T$, and any lag, h, all the first and second joint moments of $(\boldsymbol{Y}_{t_1}, ..., \boldsymbol{Y}_{t_T})'$ exists and are equal to the first and second moments of $(\boldsymbol{Y}_{t_1+h}, ..., \boldsymbol{Y}_{t_T+h})'$ [10]. In other words, a time series process, $\boldsymbol{Y}_{1:T}$, is said to be second order stationary if

1. The expected value for all the random variables in $\boldsymbol{Y_{1:T}}$ are equivalent: $E[\boldsymbol{Y}_t] = \mu$ for all t.

2. The variance for all the random variables in $\boldsymbol{Y_{1:T}}$ are equivalent: $Var[\boldsymbol{Y}_t] = \sigma$ for all t [10].

Graphical Check of Stationarity

Two diagnostic plots that are used to check the assumption of stationarity in a time series are a time series plot and a autocorrelation function (ACF) plot. A time series plot is a plot of a realization of a time series process, i.e. a plot of an observed time series. A time series plot suggest non-stationarity if it contains graphical trend and/or seasonal pattern [10].

Figure 2.2 contains four time series plots. The times series in Figures 2.2(a) and 2.2(b) were generated in R. Figure 2.2(c) is a time series plot of the data set from the R Package forecast that contains Australia's monthly gas production from 1956 to 1995 [6]. Figure 2.2(d) is a time series of the observations collected from an electroencephalogram corresponding to a patient undergoing ECT therapy [12].

Figure 2.2(a) suggest stationarity because there appears to be a constant mean and variance throughout the observed period, whereas Figure 2.2(b) suggest non-stationarity with trend because its mean increases over the observed period. Both Figures 2.2(c) and 2.2(d) suggest non-stationary with trend and seasonal pattern. The data points in each of these plots form a sinusoidal pattern and increase/decrease over the observed period.

The ACF measures the linear dependence between two random variables from the same time series process; one random variable at time t and one at time s. ACF is denoted by $\rho(t, s)$ and defined by [10]

$$\rho(t,s) \;=\; \frac{\gamma(t,s)}{\sqrt{var(\boldsymbol{Y}_t)var(\boldsymbol{Y}_s)}} \tag{2.3}$$

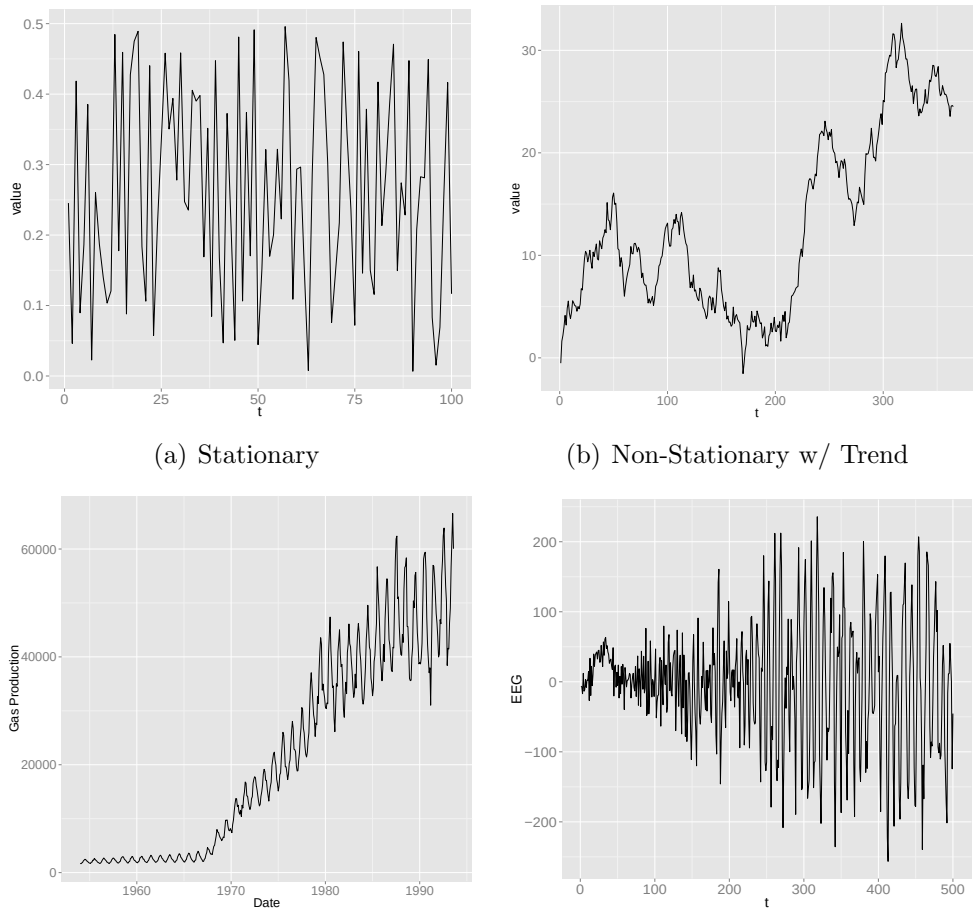
where

$\gamma(s,t) \;=\; Cov(\boldsymbol{Y}_t, \boldsymbol{Y}_s) = E\{(\boldsymbol{Y}_t \;-\; u_t)(\boldsymbol{Y}_s \;-\; u_s)\}$

$u_t \;=\; E[\boldsymbol{Y}_t]$

$u_s \;=\; E[\boldsymbol{Y}_s].$

Note that the function $\gamma(s,t)$ is the covariance function between random variables $\boldsymbol{Y_t}$ and $\boldsymbol{Y_s}$. When dealing with data that is stationary, we may assume that $E[\boldsymbol{Y}_t] = \mu$ and $Var(\boldsymbol{Y}_t) = \sigma^2$ for all values of t. Therefore, for a stationary time series process, $\gamma(t,s)$ only depends on the distance between the indices t and s, $|t-s| = h$, making $\gamma(s,t) = \gamma(h) = Cov\{\boldsymbol{Y}_t, \boldsymbol{Y}_{t-h}\}$ and $\rho(t,s) = \rho(h) = \frac{\gamma(h)}{\gamma(0)}$. An ACF plot contains the values of the autocorrelation function estimated from the time series. The ACF plot graphically explains the estimated correlation patterns displayed by a time series process at different points in time [10].



(a) Stationary

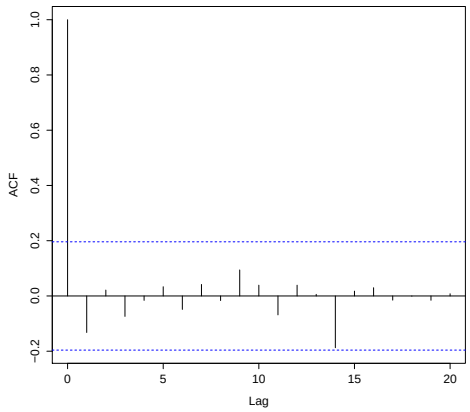
(b) Non-Stationary w/ Trend

(c) Non-Stationary w/ Trend and Seasonal Cyclic Pattern

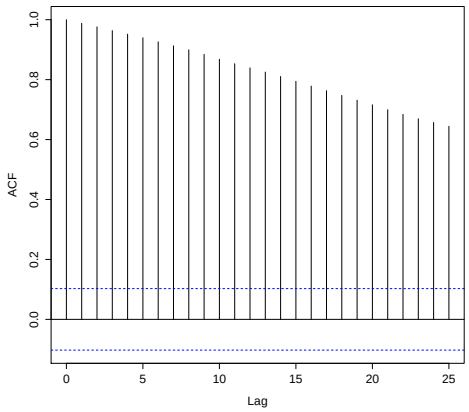
(d) Non-Stationary w/ Trend and Seasonal Cyclic Pattern

Figure 2.2: Stationary vs. Non-Stationary Time Series Plots

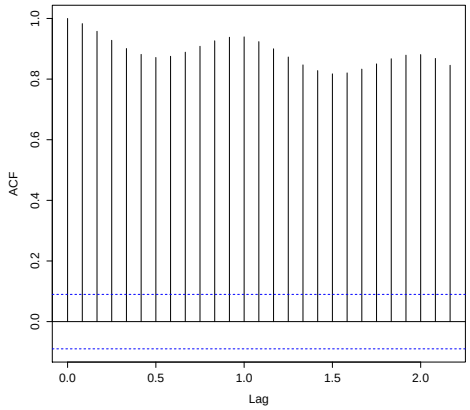
Figure 2.3 contains the corresponding ACF plots for the time series plots in Figures 2.2. The blue dotted lines in the ACF plots represent the upper and lower confidence intervals for insignificant autocorrelations. A time series that is stationary produces an ACF plot that contains values that are small and quickly approach zero as lag (h) increases. A time series that is non-stationary produces an ACF plot that contains values that are large, often positive, that slowly or sometimes never approaches zero as h increases. The values tend to lie outside of the blue dotted lines [10].



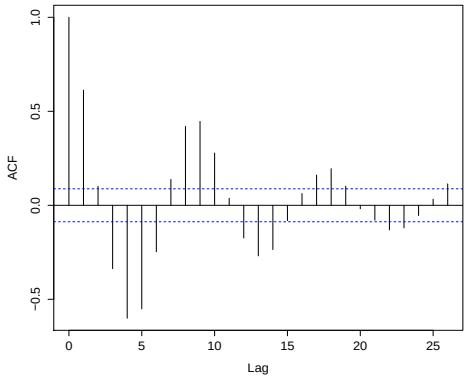
(a) Stationary



(b) Non-Stationary w/ Trend



(c) Non-Stationary w/ Trend and Seasonal Pattern



(d) Non-Stationary w/ Trend and Seasonal Pattern

Figure 2.3: Stationary vs. Non-Stationary ACF Plots

Algebraic Check of Stationarity

An algebraic method that is used to check the assumption of stationarity in a time series is the unit root test. The unit roots test that is used in this analysis is the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. The KPPS test tests for either level-stationarity or trend-stationarity. A time series process is level-stationary if it does not contain any underlying trend, whereas a time series process that is trend-stationary has underlying trend that can be removed from the time series via smoothing operators. Since we are interested in time series process that do not have any trend, we will use the level-stationarity KPSS test to test for stationarity. Thus, we will only explain the mathematics and structure of the level-stationary KPSS test in detail [8].

Let the time series $y_{1:T}$ be a realization of the time series process $\boldsymbol{Y}_{1:T}$. In order to prove that $y_{1:T}$ is a stationary time series, the KPSS test is used to check if $\boldsymbol{Y}_{1:T}$ is a stationary time series process. In order to perform the KPSS test, a linear stationary regression of $\boldsymbol{Y}_t$ on an intercept is fitted from the observed values in the time series $y_{1:T}$. The regression model is [8]

$$\boldsymbol{Y}_t \;=\; \beta_o \;+\; \boldsymbol{\varepsilon}_t \tag{2.4}$$

for $t = 1, 2, .., T$

where

$\boldsymbol{Y}_t$ = a random variable from $\boldsymbol{Y}_{1:T}$ that is indexed at time t

β_o = intercept

$\boldsymbol{\varepsilon_t}$ = random variable that represents the stationary error at time t

where

$\boldsymbol{\varepsilon}_t \overset{iid}{\sim} (0, \sigma_\varepsilon^2)$.

The null hypothesis for this test assumes that the variance of the stationary error, denoted by σ_t, is equal to zero. The alternative hypothesis assumes that the variance of the stationary error is greater than zero:

$$H_o : \sigma_\varepsilon^2 \;=\; 0 \text{ vs. } H_a : \sigma_\varepsilon^2 \;>\; 0. \tag{2.5}$$

The test statistic is denoted by $\hat{\eta}_\mu$ and defined by [8]

$$\hat{\eta_\mu} \;=\; T^{-2}\frac{\sum_{t\,=\,1}^{T}(S_t^2)}{s^2(l)} \tag{2.6}$$

where [8]

$$S_t \;=\; \sum_{i=1}^{t} e_i \tag{2.7}$$

$$s^2(l) \;=\; T^{-1}\sum_{t\,=\,1}^{T}(e_t^2) \;+\; 2T^{-1}\sum_{s\,=\,1}^{L}[(1 \;-\; \frac{s}{L \;+\; 1})\sum_{t\,=\,s+1}^{T}(e_t e_{t\,-\,s})] \tag{2.8}$$

$$L \;=\; o(T^{1/2}). \tag{2.9}$$

The residual at time is the difference between the observed and fitted value at time t. It is denoted by e_t and defined as

$$e_t \;=\; y_t \;-\; \hat{y}_t. \tag{2.10}$$

Note that $\hat{y}_t$ is the fitted value, i.e. estimated value, at time t. This estimated value is calculated from the fitted regression model, $\hat{y}_t = \hat{\beta}_o$. The fitted model only includes the estimated intercept $\hat{\beta}_o$. Therefore, $e_t = y_t - \hat{\beta}_o$ for all t. The estimated intercept is calculated from the time series $y_{1:T}$ [8].

The upper tail critical values for $\hat{\eta_\mu}$ can be found in Table 1 in Kwiatkowski et al. 1992. If $\hat{\eta_\mu}$ is greater than the critical value at the α-level, then the null hypothesis is rejected at an α-level of significance and we may assume that $y_{1:T}$ is non-stationary. If $\hat{\eta_\mu}$ is less than the critical value at the α-level, then the null hypothesis is accepted at an α-level of significance and we may assume that the time series $y_{1:T}$ is stationary [8].

It has already been shown graphically that the time series in Figure 2.2(a) is stationary. The KPSS test will be used to algebraically show that the time series in Figure 2.2(a) is stationary. The value of the test statistic that is calculated from this time series is $\hat{\eta_\mu} = 0.0390$. In Kwiatkowski et al. 1992, Table 1 shows that at the $\alpha = .05$ level of significance, the p-value for the given value of $\hat{\eta_\mu}$ is .10. This implies that we do not have enough evidence to reject the null hypothesis and that this time series has unit root. Therefore, the time series in Figure 2.2(a) has been shown to be graphically and algebraically stationary.

Removing Non-Stationarity

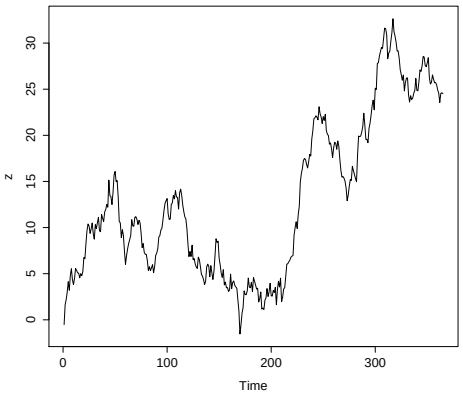
In time series analysis, the assumption of stationarity is important when modeling a time series. Because most time series are realizations of a non-stationary process, there exists smoothing techniques that are applied to time series in order to separate the non-stationary data from the stationary data. These techniques involve decomposing the time series into a "smooth" component and a component that contains the unexplained white noise. Two smoothing techniques that are widely used and that are used in this analysis are moving averages and differencing [10].

Suppose $y_{1:T}$ is a non-stationary time series. A technique that is used to remove the non-stationarity from $y_{1:T}$ is by applying the moving average operator M, to $y_{1:T}$. Applying M to $y_{1:T}$ returns the set of transformed values $\{z_{q+1}, z_{q+1}, ..., z_{T-p}\}$ where [10]

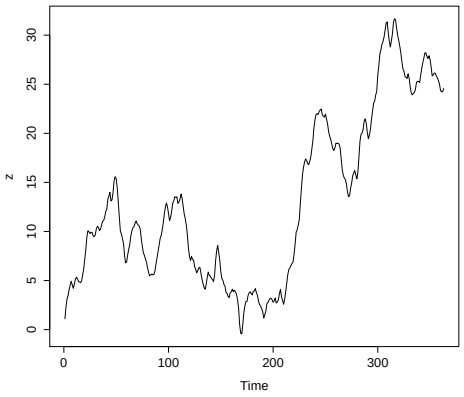
$$z_t \ = \ \sum_{j \, = \, -q}^{p} (a_j y_{j \, + \, j}), \ \text{for } t = (q \ + \ 1) : (T \ - \ p), \tag{2.11}$$

where a_j's are weights that sum to one, $a_j \geq 0$ for all j and $a_j = a_{-j}$. It is generally assumed that $p = q$, p and q are small values and an equal weight is selected for all j. The order of M is equal to $2p + 1$ [10].

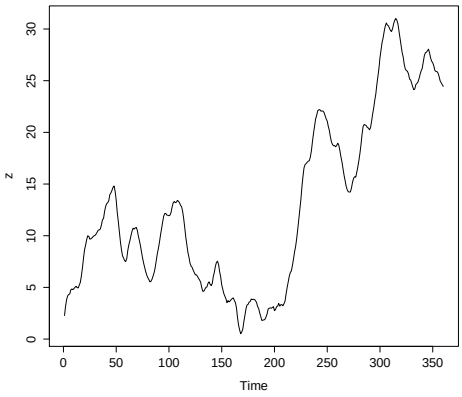
Figure 2.4 contains the results of applying the moving average operator with orders $q = 3$ and $q = 6$ to the non-stationary time series in Figure 2.2. Figures 2.4(b) and 2.4(c) show that the transformed values in the set $\{z_{q+1}, z_{q+1}, ..., z_{T-p}\}$ become smoother as the order of the moving average operator increases. The moving average smoothing technique removes the white nose from the time series, but it does not remove the increasing trend.



(a) Non-Stationary w/ Trend



(b) Moving Average $q = 3$



(c) Moving Average $q = 6$

Figure 2.4: Graphical results of applying the moving average operator to a non-stationary time series with increasing trend

There exists two types of differencing techniques, non-seasonal differencing and seasonal differencing. Non-seasonal differencing is a smoothing technique that removes trend from a time series. The d-order non-seasonal differencing operator is denoted by D^d and defined by [9]

$$D^d \;=\; (1 \;-\; B)^d. \tag{2.12}$$

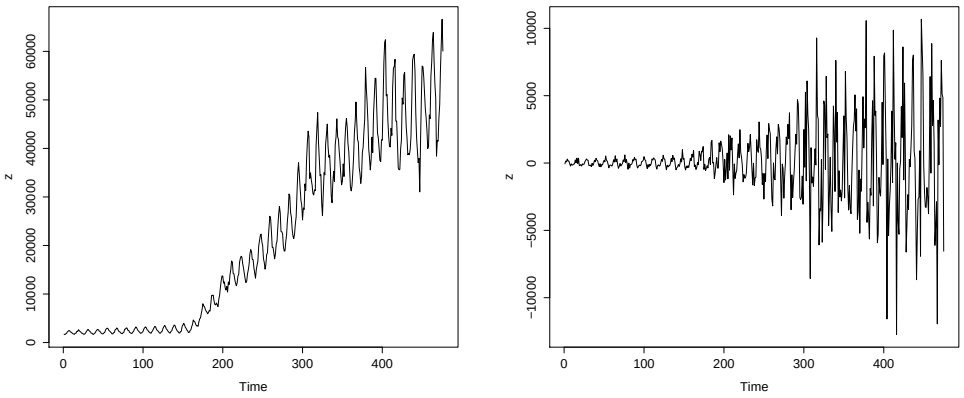
D^d is a function of the back-shift operator. The back-shift operator is denoted by B and is defined by

$$By_t \;=\; y_{t-1}. \tag{2.13}$$

Applying the first-order non-seasonal differencing operator (D^1) to the time series $y_{1:T}$, re-

turns the set of transformed values $\{z_1, .., z_{T-1}\}$, where [9]

$$z_t \ = \ D^1 y_t \ = \ (1 - B)^1 y_t \ = \ y_t \ - \ B y_t \ = \ y_t \ - \ y_{t-1} \ []. \tag{2.14}$$



(a) Australian monthly gas production time series plot.

(b) Time series plot of the D^1 transformed values $\{z_1, .., z_{T-1}\}$.

Figure 2.5: Graphical results of applying the operator D^1 to the Australian monthly gas production time series.

Figure 2.5(a) is the time series plot of the Australian monthly gas production. It was mentioned previously that it is an example of a non-stationary time series with increasing trend and a seasonal pattern. Figure 2.5(b) is a time series plot of the results of applying D^1 to the Australian monthly gas production time series. The time series plot in Figure 2.5(b) does no appear to have an increasing trend. However, there still appears to be a seasonal pattern.

The seasonal differencing technique is a smoothing technique that removes both trend and seasonal patterns from a non-stationary time series. The s-period d-order seasonal differencing operator is denoted by D_s^d. It is defined by [9]

$$D_s^d \ = \ (1 \ - \ B^s)^d \tag{2.15}$$

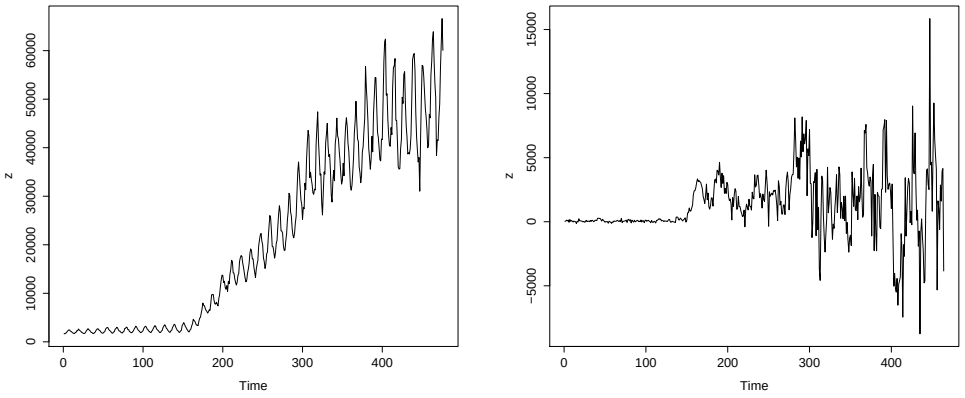
and it is a function of the s-degree back shift operator B^s which is defined by

$$B^s y_t \ = \ y_t \ - \ y_{t-s}. \tag{2.16}$$

It is also a function of the number of periods in a season, denoted by s. The data points in the Australian monthly gas production were collected every month. Therefore, in order to

remove the seasonal pattern from the time series, the 12-period (monthly) first-order seasonal difference operator, D_{12}^1, is applied to the Australian gas production time series, $y_{1:T}$. This returns a vector of transformed values $\{z_1, .., z_{T-1}\}$ where [9]

$$z_t \;=\; D_{12}^1 y_t \;=\; (1 - B^{12})^1 y_t \;=\; y_t - B^{12} y_t \;=\; y_t - y_{t-12}\ [9]. \tag{2.17}$$



(a) Australian monthly gas production time series plot.

(b) Time series plot of the D_{12}^1transformed values $\{z_1, .., z_{T-1}\}$.

Figure 2.6: Graphical results of applying D_{12}^1 to the Australian monthly gas production time series.

Figure 2.6(b) is a time series plot that contains the results of applying D_{12}^1 to the Australian monthly gas production time series. There no longer appears to be an increasing trend nor seasonal pattern, implying that the D_{12}^1 operator removed the non-stationarity from the Australian monthly gas production time series.

Chapter 3

Extreme Values in a Time Series Process

Extreme Value Theory - Stationary

Estimating the distribution of an event that occurs with very small probability is of interest in statistical inference. In statistics, these rare events are referred to as extreme events and the theory of modeling and measuring these events is known as Extreme Value Theory (EVT). Examples of statistics that are extreme events are maximums and minimums. Order Statistics is a statistical method that infers on the distributions of these extreme events [4].

Order Statistics

Let $\boldsymbol{X}_{1:n} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n\}$ be an random sample. The random variables in $\boldsymbol{X}_{1:n}$ placed in ascending order are known as the order statistics of $\boldsymbol{X}_{1:n}$. They are denoted by $\boldsymbol{X}_{(i:n)} = \{\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}, ..., \boldsymbol{X}_{(n)}\}$ and they satisfy $\boldsymbol{X}_{(1)} \leq \boldsymbol{X}_{(2)} \leq ... \leq \boldsymbol{X}_{(n)}$. The order statistics are defined as

$$\begin{aligned}
\boldsymbol{X}_{(1)} &= min\{\boldsymbol{X}_{1:n}\} \\
\boldsymbol{X}_{(2)} &= \ second\ smallest\{\boldsymbol{X}_{1:n}\} \\
&\vdots \\
\boldsymbol{X}_{(n-1)} &= \ second\ largest\{\boldsymbol{X}_{1:n}\} \\
\boldsymbol{X}_{(n)} &= max\{\boldsymbol{X}_{1:n}\}.
\end{aligned}$$

Since order statistics are random variables, each order statistic has a unique pdf that is calculated from the joint probability distribution of the random sample $\boldsymbol{X}_{1:n}$ [2].

Let $\boldsymbol{Z}_{1:m} = \{\boldsymbol{Z}_1, ..\boldsymbol{Z}_m\}$ be a set of independent random variables where $f_{\boldsymbol{Z}_i}(z_i)$ is the pdf for the random variable $\boldsymbol{Z}_i$. The joint probability distribution function of a set of independent random variables is equal to the product of each of the random variables' pdfs. Therefore, the joint pdf for the set $\boldsymbol{Z}_{1:m}$ is defined by [2]

$$f_{\boldsymbol{Z}_1,...,\boldsymbol{Z}_m}(\boldsymbol{Z}_1, ..., \boldsymbol{Z}_m) = \prod_{i=1}^{m} f_{\boldsymbol{Z}_i}(z_i) = f_{\boldsymbol{Z}_1}(z_1) * ... * f_{\boldsymbol{Z}_m}(z_m). \tag{3.1}$$

As mentioned previously, a random sample is a set of independent random variables that have the same pdf. Therefore, the joint pdf of the random sample $\boldsymbol{X}_{1:n}$ is defined by

$$f_{\boldsymbol{X}_1,...,\boldsymbol{X}_n}(x_1, ..., x_n) = \prod_{i=1}^{n} f_{\boldsymbol{X}}(x) = [f_{\boldsymbol{X}}(x)]^n \tag{3.2}$$

where $f_{\boldsymbol{X}}(x)$ is the pdf for the random variables in $\boldsymbol{X_{1:n}}$ [2]. Unlike a set of independent random variables or, more specifically, a random sample, the joint pdf of a set of random variables that are dependent is very difficult to calculate. Since the random variables in a time series process are dependent, the joint probability distribution of a time series process is almost impossible to calculate. Therefore, modeling the distribution of extreme values in a time series process can not be done through order statistics.

Generalized Extreme Value Distribution

Let $\boldsymbol{Z}_{1:n} = \{\boldsymbol{Z}_1, , ..., \boldsymbol{Z}_n\}$ be a set of independent random variables. In Extreme Value Theory, it has been shown that as $n \to \infty$, the Generalized Extreme Value Distribution (GEV) is the limiting distribution of the order statistics $\boldsymbol{Z}_{(min)}$ and $\boldsymbol{Z}_{(max)}$ [4]. The GEV is a family of pdfs that combine the Gumbel, Frechet and Weibull distributions. The pdf for $\boldsymbol{Z}_{(max)}$ is defined by [4]

$$f_{\boldsymbol{Z}_{(max)}}(z|\mu, \sigma, \xi) = \frac{1}{\sigma} t(z)^{\xi+1} e^{-t(z)} \tag{3.3}$$

where

$$t(z) = \begin{cases} (1 + (\frac{z-\mu}{\sigma}\xi))^{\frac{1}{\xi}}, & \text{if } \xi \neq 0 \\ e^{-\frac{z-\mu}{\sigma}}, & \text{if } \xi = 0 \end{cases} \tag{3.4}$$

and where

μ = location parameter

σ = scale parameter (> 0)

ξ = shape parameter.

The pdf for $\boldsymbol{Z}_{(min)}$ is denoted by $f_{\boldsymbol{Z}_{(min)}}(z|\mu,\sigma,\xi)$. Since it can be shown that $min(\boldsymbol{Z}_1,...,\boldsymbol{Z}_n) = -max(-\boldsymbol{Z}_1,...,-\boldsymbol{Z}_n)$, $f_{\boldsymbol{Z}_{(min)}}(z|\mu,\sigma,\xi)$ is defined in terms of $f_{\boldsymbol{Z}_{(max)}}(z|\mu,\sigma,\xi)$ where

$$f_{\boldsymbol{Z}_{(min)}}(\boldsymbol{Z}|\mu,\sigma,\xi) = 1 - f_{\boldsymbol{Z}_{(max)}}(\boldsymbol{Z}|\mu,\sigma,\xi). \tag{3.5}$$

Let $\boldsymbol{X}_{1:n}$ be a random sample. Since a random sample is set of independent random variables, for large values of n, the pdfs for $\boldsymbol{X}_{(min)}$ and $\boldsymbol{X}_{(max)}$ are denoted by $f_{\boldsymbol{X}_{(min)}}(x|\mu,\sigma,\xi)$ and $f_{\boldsymbol{X}_{(max)}}(x|\mu,\sigma,\xi)$ and are equivalent to equations 3.3 and 3.5, respectively,

$$f_{\boldsymbol{X}_{(min)}}(x|\mu,\sigma,\xi) = f_{\boldsymbol{Z}_{(min)}}(z|\mu,\sigma,\xi) \tag{3.6}$$

$$f_{\boldsymbol{X}_{(max)}}(x|\mu,\sigma,\xi) = f_{\boldsymbol{Z}_{(max)}}(z|\mu,\sigma,\xi). \tag{3.7}$$

Extreme Value Theorem for Stationary Time Series Processes

Let $\boldsymbol{Y}_{1:n} = \{\boldsymbol{Y}_1,..,\boldsymbol{Y}_n\}$ be a stationary time series process and let

$$\boldsymbol{Y}_{(min)} = min\{\boldsymbol{Y}_1,...,\boldsymbol{Y}_n\} \quad \text{and} \quad \boldsymbol{Y}_{(max)} = min\{\boldsymbol{Y}_1,...,\boldsymbol{Y}_n\}.$$

Unlike a random sample, the random variables in $\boldsymbol{Y}_{1:n}$ are neither independent nor identically distributed. Since neither of the assumptions of independence or an identical distribution are met, it would seem that the limiting distributions for $\boldsymbol{Y}_{(min)}$ and $\boldsymbol{Y}_{(max)}$ are not from the family of generalized extreme value distributions. However, the Extreme Value Theorem for Stationary Time Series Process (EVTS) states that for large values of n, $\boldsymbol{Y}_{(min)}$ and $\boldsymbol{Y}_{(max)}$ follow a GEV distribution when $\boldsymbol{Y}_{(min)}$ and $\boldsymbol{Y}_{(max)}$ are from a stationary time series process. In particular, the pdfs of $\boldsymbol{Y}_{(min)}$ and $\boldsymbol{Y}_{(max)}$ are denoted by $f_{\boldsymbol{Y}_{(min)}}(y|\mu,\sigma,\xi)$ and $f_{\boldsymbol{Y}_{(max)}}(y|\mu,\sigma,\xi)$, and are also equivalent to equations 3.3 and 3.5, respectively,

$$f_{\boldsymbol{Y}_{(min)}}(y|\mu,\sigma,\xi) = f_{\boldsymbol{Z}_{(min)}}(z|\mu,\sigma,\xi) \tag{3.8}$$

$$f_{\boldsymbol{Y}_{(max)}}(y|\mu,\sigma,\xi) = f_{\boldsymbol{Z}_{(max)}}(z|\mu,\sigma,\xi). \tag{3.9}$$

A full explanation and proof of the EVTS can be found in Haan et. al 2006.

The Block Minima/Maxima Method and the Peaks Over Threshold Method are two techniques that are used to estimate the parameters in $f_{\boldsymbol{Y}_{(min)}}(y|\mu,\sigma,\xi)$ and $f_{\boldsymbol{Y}_{(max)}}(y|\mu,\sigma,\xi)$. Both of these techniques are based off of the EVTS. For this analysis, we will only explain the Block Minima Method because it is used to estimate the distribution of an application's minimum daily transactional response time.

Block Minima Method

Let $y_{1:n}$ be a time series from the stationary time series process $\boldsymbol{Y}_{1:n}$. In extreme value theory, the block minima method consists of dividing the observation in $y_{1:T}$ into K non-overlapping blocks of equal size [3]. The set $y_{(min)_{1:K}} = \{m_1, m_2, ..., m_K\}$ consists of the observed minimum values from each of the K blocks. The observations in $y_{(min)_{1:K}}$ are defined by

$$\begin{aligned} m_1 &= min(\text{block 1}) \\ m_2 &= min(\text{block 2}) \\ &\vdots \\ m_k &= min(\text{block } K). \end{aligned}$$

In the EVTS, it has been proven that the limiting distribution of $\boldsymbol{Y}_{(min)}$ (eqn. 3.8) can be estimated by the pdf $f_{\boldsymbol{Y}_{(min)}}(y|\hat{\mu},\hat{\sigma},\hat{\xi})$, which is defined by

$$f_{\boldsymbol{Y}_{(min)}}(y|\hat{\mu},\hat{\sigma},\hat{\xi}) \;=\; 1 - f_{\boldsymbol{Y}_{(max)}}(-y|\hat{\mu},\hat{\sigma},\hat{\xi}) \;=\; 1 - \frac{1}{\hat{\sigma}} t(-y)^{\hat{\xi}+1} e^{-t(-y)} \tag{3.10}$$

where

$$t(-y) = \begin{cases} (1 + (\frac{-y-\hat{\mu}}{\hat{\sigma}}\hat{\xi}))^{\frac{1}{\hat{\xi}}}, & \text{if } \hat{\xi} \neq 0 \\ e^{-\frac{-y-\hat{\mu}}{\hat{\sigma}}}, & \text{if } \hat{\xi} = 0 \end{cases}$$

and where

$\hat{\mu}$ = estimated location parameter

$\hat{\sigma}$ = estimated scale parameter (> 0)

$\hat{\xi}$ = estimated shape parameter.

The estimated parameters, $\hat{\mu}$, and $\hat{\sigma}$, $\hat{\xi}$ are calculated directly from the observations in $y_{(min)_{1:K}}$ [3].

Estimating Parameters

A method used to estimate the parameters in $f_{\boldsymbol{Y}_{(min)}}(y|\hat{\mu},\hat{\sigma},\hat{\xi})$ is the maximum likelihood method. The maximum likelihood estimates of the parameters are the values which the likelihood function, denoted by $L(\mu,\sigma,\xi|y_{(min)_{1:K}})$, attains its maximum. An in depth explanation of this method can be found in Casella et. al. 2002 [2]. Once the parameters are estimated via the maximum likelihood method, $E[\{\boldsymbol{Y}_{(min)}]$ is estimated directly from the estimated parameters $\hat{\mu}$, and $\hat{\sigma}$, $\hat{\xi}$ and is defined by [3]

$$\hat{E}[\{y_{(min)}\}] = -[\hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} + \frac{\hat{\sigma}}{\hat{\xi}}g_1], \tag{3.11}$$

where $g_k = \Gamma(1 - k\hat{\xi})$ and $\Gamma(t)$ is the gamma function. The variance of $\boldsymbol{Y}_{(min)}$ is also estimated from the estimated parameters and is defined by [3]

$$\hat{Var}[\{y_{(min)}\}] = \frac{\sigma^2}{\xi^2}(g_2 - g_1)^2. \tag{3.12}$$

The estimated pdf $f_{\boldsymbol{Y}_{(min)}}(y|\hat{\mu},\hat{\sigma},\hat{\xi})$ is beneficial because it is related to the estimated cumulative distribution function (cdf). The estimated cdf is denoted as $F_{\boldsymbol{Y}_{(min)}}(z|\hat{\mu},\hat{\sigma},\hat{\xi})$ and defined by

$$F_{\boldsymbol{Y}_{(min)}}(z|\hat{\mu},\hat{\sigma},\hat{\xi}) = P(\boldsymbol{Y}_{(min)} \leq z|\hat{\mu},\hat{\sigma},\hat{\xi}) = \int_0^z f_{\boldsymbol{Y}_{(min)}}(y|\hat{\mu},\hat{\sigma},\hat{\xi})\, dy. \tag{3.13}$$

The cdf calculates the probability of $\boldsymbol{Y}_{(min)}$ being less than or equal to z. The cdf is also used to construct the lower and upper limits for confidence intervals. For example, suppose we wanted to calculate the lower and upper limits for a 95% confidence interval. The lower and upper limits are denoted by q_{lower} and q_{upper} and satisfy equations 3.15 and 3.14, respectively,

$$F_{\boldsymbol{Y}_{(min)}}(q_{lower}|\hat{\mu},\hat{\sigma},\hat{\xi}) = P(\boldsymbol{Y}_{(min)} \leq q_{lower}) = \int_0^{q_{lower}} f_{\boldsymbol{Y}_{(min)}}(y|\hat{\mu},\hat{\sigma},\hat{\xi})\, dy = .025 \tag{3.14}$$

$$F_{\boldsymbol{Y}_{(min)}}(q_{upper}|\hat{\mu},\hat{\sigma},\hat{\xi}) = P(\boldsymbol{Y}_{(min)} \leq q_{upper}) = \int_0^{q_{upper}} f_{\boldsymbol{Y}_{(min)}}(y|\hat{\mu},\hat{\sigma},\hat{\xi})\, dy = .975. \tag{3.15}$$

Statistics that perform well with data that is drawn from a wide range of pdfs are known as robust statistics. Since the expected value of a random variable is a weighted average, it is sometimes considered to be a non-robust statistic because it can be effected by outliers. The median is a statistic that is often used in place of the expected value because it is a robust statistic. The median is the value that separates the higher half of a probability distribution from the lower half [2]. The estimated median of $\boldsymbol{Y}_{(min)}$ is denoted by $q_{.50}$. It too is calculated directly from the estimated parameters and defined by [3]

$$q_{.50} = \begin{cases} -[\hat{\mu} + \hat{\sigma}\frac{(ln2)^{-\hat{\xi}}-1}{\hat{\xi}}], & \text{if } \hat{\xi} \neq 0 \\ -[\hat{\mu} - \hat{\sigma} ln[ln(2)]], & \text{if } \hat{\xi} = 0. \end{cases} \tag{3.16}$$

The median can also be the estimated through the cdf, where $q_{.50}$ satisfies

$$F_{\boldsymbol{Y}_{(min)}}(q_{.50}|\hat{\mu}, \hat{\sigma}, \hat{\xi}) = P(\boldsymbol{Y}_{(min)} \leq q_{.50}) = P(\boldsymbol{Y}_{(min)} \geq q_{.50}) = .50. \tag{3.17}$$

Extreme Value Theory - Non-Stationary

In the previous section, we showed that with the EVTS, the block minima method and the assumption of stationarity, $f_{\boldsymbol{Y}_{(min)}}(y|\hat{\mu}, \hat{\sigma}, \hat{\xi})$ is the estimated pdf for the minimum of a time series process (eqn. 3.10). We have discussed in earlier sections that time series are rarely stationary. Since the assumption of stationarity is crucial in the EVTS and the block minima method, the non-stationary generalized extreme value (non-stationary GEV) distribution was developed in order to model the distribution of the minimum of a non-stationary time series process. The GEV distribution defined earlier in eqn. 3.3 will now be referred to as the stationary GEV distribution.

Non-Stationary Generalized Extreme Value Distribution

In a stationary time series, the mean and variance are constant over the observation period. Because of this, the parameters in the stationary-GEV distribution are fixed. The non-stationary-GEV distribution accounts for trend and/or seasonal pattern by making one or more of the model parameters functions of time (t) and/or seasonal period (s). The time/seasonal dependent parameters are denoted as $\mu(t,s), \sigma(t,s),$ and $\xi(t,s)$. The interpretation of these parameters are identical to their interpretation in the stationary-GEV distribution. However, since the parameters $\mu(t,s), \sigma(t,s),$ and $\xi(t,s)$ are functions of t and s, they are not constant.

If a realization of a time series appears to only have trend, one or more of the parameters are typically polynomials that are dependent on t. If a time series appears to only have

seasonal pattern, one or more of the parameters are typically sinusoidal functions that are dependent on s. If a time series appears to have trend and seasonal pattern, one or more of the parameters are typically functions that are dependent on t and s. Table 3.1 contains the structure of the three types of non-constant parameter functions.

Table 3.1: Time dependent parameter functions for the non-stationary GEV pdf.

Time (t)	Seasonal (s)	Time and Seasonal (t, s)
$\mu(t) = \mu_o + \sum_{k=1}^{p_\mu} \mu_k t^k,$	$\mu(s) = \mu_o + \mu_{sin} sin(\omega c_s) + \mu_{cos} cos(\omega c_s)$	$\mu(t,s) = \mu(t) + \mu(s)$
$\sigma(t) = \sigma_o + \sum_{k=1}^{p_\sigma} \sigma_k t^k$	$\sigma(s) = \sigma_o + \sigma_{sin} sin(\omega c_s) + \sigma_{cos} cos(\omega c_s)$	$\sigma(t,s) = \sigma(t) + \sigma(s)$
$\xi(t) = \xi_o + \sum_{k=1}^{p_\xi} \xi_k t^k$	$\xi(s) = \xi_o + \xi_{sin} sin(\omega c_s) + \xi_{cos} cos(\omega c_s)$	$\xi(t,s) = \xi(t) + \xi(s)$

Note: The set of parameters $\{p_\mu, p_\sigma, p_\xi\}$ represents the degrees of their corresponding polynomial functions. The sets of parameters $\theta_\mu = \{\mu_o, \mu_1, ..., \mu_{p_\mu}, \mu_{sin}, \mu_{cos}\}$, $\theta_\sigma = \{\sigma_o, \sigma_1, ..., \sigma_{p_\sigma}, \sigma_{sin}, \sigma_{cos}\}$ and $\theta_\xi = \{\xi_o, \xi_1, ..., \xi_{p_\xi}, \xi_{sin}, \xi_{cos}\}$ are the coefficients for their corresponding functions. The parameter $\omega = \frac{2\pi}{365.25}$ and the variable c_s denotes the center of the s-th period counted in days starting from the beginning of the year.

Model Selection

Let $y_{1:T}$ be a time series that is a realization of the time series process $\boldsymbol{Y}_{1:T}$. Suppose we are interested in estimating the parameters in $f_{\boldsymbol{Y}_{(min)}}(y|\hat{\mu}, \hat{\sigma}, \hat{\xi})$ from the time series $y_{1:T}$. The first step in estimating the the parameters is determining if $y_{1:T}$ is stationary. The stationarity of a time series is visually analyzed through the time series and ACF plots of $y_{1:T}$. The stationarity of $y_{1:T}$ is then analyzed algebraically with the KPSS unit root test.

If the results of these plots and test suggest that $y_{1:T}$ is a stationary time series, then the block minima method is implemented and the parameters for the stationary-GEV distribution are estimated from the set of block minimums $y_{(min)_{1:K}}$. However, if the results suggest that $y_{1:T}$ is a non-stationary time series, then the block minima method is still implemented but the parameters for the non-stationary GEV distribution are estimated from $y_{(min)_{1:K}}$.

In R, there exists functions that only estimate the parameters in $f_{\boldsymbol{Y}_{(max)}}(y|\mu,\sigma,\xi)$. Therefore, a problem arises in R when estimating the parameters in $f_{\boldsymbol{Y}_{(min)}}(y|\mu,\sigma,\xi)$. Equation 3.10 states that $f_{\boldsymbol{Y}_{(min)}}(y|\mu,\sigma,\xi)$ is defined as

$$f_{\boldsymbol{Y}_{(min)}}(y|\hat{\mu},\hat{\sigma},\hat{\xi}) = 1 - f_{\boldsymbol{Y}_{(max)}}(y|\hat{\mu},\hat{\sigma},\hat{\xi}).$$

Therefore, fitting the negative of the block minimums $(-y_{(min)_{1:K}})$ to $f_{\boldsymbol{Y}_{(max)}}(y|\mu,\sigma,\xi)$ would avoid this problem in R. However, the functions in R that would be used to perform this fit have a difficult time handling data with negative values.

It can be proven that $\frac{1}{min(\boldsymbol{Y}_1,...,\boldsymbol{Y}_n)} = max(\frac{1}{\boldsymbol{Y}_1},...,\frac{1}{\boldsymbol{Y}_n})$. Therefore, the following transformation and analysis will be performed in R to to estimate the mean, median and confidence bands for $\boldsymbol{Y_{(min)}}$:

1. Perform a normalized reciprocal transformation on the set of block minimums $(y_{(min)_{1:K}})$. The purpose of normalizing the data is that it returns a better fit and a lower AIC value. The set of the transformed block minimums is denoted as $z_{(min)_{1:K}} = \{z_1, z_2, .., z_K\}$. The transformed minimum for block k is denoted as z_k and defined by

$$z_k = \frac{\frac{1}{m_k} - \mu_{recip}}{v_{recip}} \tag{3.18}$$

   where

   $m_k = min(block_k)$

   $\mu_{recip} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m_k}$

   $v_{recip} = \frac{1}{K}\sum_{k=1}^{K}(\frac{1}{m_k} - \mu_{recip})^2$.

2. Fit the transformed values $z_{(min)_{1:K}} = \{z_1, z_2, .., z_K\}$ to $f_{\boldsymbol{Y}_{(max)}}(y|\mu,\sigma,\xi)$ in R with the function $gev.fit$ from the ismev package [5].

3. Once the estimated parameters are calculated, the estimated expected value and median of $\boldsymbol{Y_{(min)}}$ are calculated from the following two equations:

$$\hat{E}[\{y_{(min)}\}] = [\hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} + \frac{\hat{\sigma}}{\hat{\xi}}g_1] * v_{recip} + \mu_{recip}, \tag{3.19}$$

$$q_{.50} = \begin{cases} [\hat{\mu} + \hat{\sigma}\frac{(ln2)^{-\hat{\xi}}-1}{\hat{\xi}}] * v_{recip} + \mu_{recip}, & \text{if } \hat{\xi} \neq 0 \\ [\hat{\mu} - \hat{\sigma}ln[ln(2)]] * v_{recip} + \mu_{recip}, & \text{if } \hat{\xi} = 0. \end{cases} \tag{3.20}$$

The lower and upper 95% confidence bands are denoted as q^*_{lower} and q^*_{upper}, respectively, and are calculated by

$$q^*_{lower} = \frac{1}{q_{lower} * v_{recip} + \mu_{recip}} \text{ and } q^*_{upper} = \frac{1}{q_{upper} * v_{recip} + \mu_{recip}} \tag{3.21}$$

where

q_{lower} satisfies $F_{\boldsymbol{Y}_{(max)}}(q_{lower}|\hat{\mu}, \hat{\sigma}, \hat{\xi}) = .025$. and q_{upper} satisfies $F_{\boldsymbol{Y}_{(max)}}(q_{upper}|\hat{\mu}, \hat{\sigma}, \hat{\xi}) = .975$.

In Step 2, the technique that the function $gev.fit()$ uses to estimate the parameters for either the stationary or non-stationary GEV distribution is the Nelder-Mead Method. The Nelder-Mead Method is a non linear optimization method that finds the maximum likelihood estimates for the parameters [3].

Chapter 4

Modeling Time Series

Introduction

In order to understand the relationship between the sequential observations in a time series and be able to forecast future observations, time series are fitted to statistical models. For example, recall the time series, $y_{1:365}$, that is plotted in Figure 2.1 contains an application's daily average response times. Questions that may arise when analyzing this time series is how a day's average response time is related to the previous days' average response times or if the day of the week or month has influence on a day's average response time. Fitting $y_{1:365}$ to an appropriate statistical model will provide answers to these questions.

There are different types of statistical models that can be used to analyze a time series. The selection of the model often depends on the stationarity of the time series. If a time series is stationary, either an autoregressive, a moving average, or an autoregressive moving average model can be selected as the most appropriate model. If the time series is non-stationary, a mixed model that incorporates smoothing techniques is often selected as the most appropriate model. Two non-stationary mixed models that are discussed in this analysis are the seasonal and non-seasonal autoregressive integrated moving average models. Again, the selection of the most appropriate mixed model depends on the type of non-stationarity that is present in the time series.

Non-Mixed Models

Autoregressive Model: $AR(p)$

If the time series $y_{1:365}$ is a realization of the stationary time series process that assumes a day's average response time is dependent only on the previous days' average response times, then an autoregressive model with order p ($Ar(p)$) would be used to model the dependencies in $y_{1:365}$. The $AR(p)$ model is the simplest time series model that is used to explore the dependencies between random variables in a stationary time series process. The order of an

autoregressive time series model is denoted by p and it represents the maximum distance ($|t-s|$) between random variables $\boldsymbol{Y_t}$ and $\boldsymbol{Y_s}$ such that $cov(\boldsymbol{Y_t}, \boldsymbol{Y_s})$ is significant [10].

The time series process $\boldsymbol{Y_{1:T}}$ is an $Ar(p)$ process if each $\boldsymbol{Y_t}$ arises from the autoregressive time series model [10]

$$\boldsymbol{Y_t} = \sum_{j=1}^{p} \phi_j \boldsymbol{Y_{t-j}} + \varepsilon_t \tag{4.1}$$

where

p = model order

ϕ_j = model parameter for $\boldsymbol{Y_{t-j}}$

ε_t = stationary error at time t.

It is often assumed that for all t, $\varepsilon_t \overset{iid}{\sim} normal(0, v)$. Suppose $\boldsymbol{Y_t}$ arises from an autoregressive time series process with order p = 3. This implies that the value of $\boldsymbol{Y_t}$ depends on $\boldsymbol{Y_{t-1}}$,$\boldsymbol{Y_{t-2}}$ and $\boldsymbol{Y_{t-3}}$ and the stationary error at time t and is equivalent to

$$\boldsymbol{Y_t} = \sum_{j=1}^{3} \phi_j \boldsymbol{Y_{t-j}} + \varepsilon_t, = \phi_1 \boldsymbol{Y_{t-1}} + \phi_2 \boldsymbol{Y_{t-2}} + \phi_3 \boldsymbol{Y_{t-3}} + \varepsilon_t.$$

The autoregressive characteristic polynomial of an $AR(p)$ process is denoted as Φ and defined by [10]

$$\Phi = 1 - \sum_{j=1}^{p} \phi_j u^j \text{ [10].} \tag{4.2}$$

If $\Phi(u) = 0$ only for values of u such that $|u| < 1$, then the autoregressive characteristic polynomial, Φ, has unit root. Causality is the direct cause and effect relationship between two observations. An AR(p) process is *casual* if Φ has unit root. Causality of an $AR(p)$ process implies stationarity, however, stationarity does not imply causality [10].

When fitting $y_{1:T}$ to an $AR(p)$ model, the set of model parameters $\{\phi\} = \{\phi_1.\phi_2, ..., \phi_p\}$ are estimated directly from $y_{1:T}$. The set of estimated parameters are denoted by $\{\hat{\phi}\}$ $= \{\hat{\phi}_1, \hat{\phi}_2, ..., \hat{\phi}_p\}$. The most common technique that is used to estimate the parameters in a statistical model is the maximum likelihood method [10]. The maximum likelihood

estimates of the parameters $\{\phi\}$, are the values which the likelihood function, denoted by $L[\{\phi\}|\{y_{1:T)}\}]$, attains its maximum [2].

The appropriate order of an $AR(p)$ model is chosen via the Akaike Information Criterion (AIC) method. The AIC method involves fitting $y_{1:T}$ to AR models for different values of p. The value of p that returns the fitted AR model with the lowest AIC value is then selected as the most appropriate model. The equation for AIC is [10]

$$AIC = 2k - 2ln(L) \tag{4.3}$$

where

k = number of parameters in the model

$L = L[\{\hat{\phi}\}|\{y_{1:T)}\}]$.

Once $\{\hat{\phi}\}$ are estimated and p is selected, the fitted values of $y_{1:T}$, denoted by $\hat{y}_{1:T}$, are estimated from the fitted $AR(p)$ model

$$\hat{y}_t = \sum_{j=1}^{p} \hat{\phi}_j \hat{y}_{t-j} \tag{4.4}$$

where

$\hat{y}_1 = y_1,\ \hat{y}_2 = y_2,\ ...,\ \hat{y}_p = y_p$.

Moving Average Model: $MA(q)$

If the time series $y_{1:365}$ is a realization of the stationary time series process that assumes a day's average response time is dependent only on the previous days' average response times' random errors, then a moving average time series model with order q ($MA(q)$) would be used to model the dependencies in $y_{1:365}$. The $MA(q)$ model is used to model the random error dependencies in a time series process. The parameter q represents the order of a moving average model and it is defined as the maximum distance ($|t - s|$) between the random variables $\boldsymbol{Y_t}$ and $\boldsymbol{Y_s}$ such that $cov(\varepsilon_t,\varepsilon_s)$ is significant [10].

The time series process $\boldsymbol{Y_{1:T}}$ is an $MA(q)$ process if each $\boldsymbol{Y_t}$ arises from the moving average time series model [10]

$$\boldsymbol{Y_t} = \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t, \tag{4.5}$$

where

q = model order

θ_j = model parameters for ε_{t-j}

ε_t = stationary error at time t.

It is often assumed that $\varepsilon_t \overset{iid}{\sim} normal(0, v)$. The moving average characteristic polynomial of an $MA(q)$ process is denoted by Θ and defined by [10]

$$\Theta = 1 - \sum_{j=1}^{q} \theta_j u^j. \tag{4.6}$$

If $\Theta(u) = 0$ only for $|u| < 1$, then Θ has unit root. An important assumption for a $MA(q)$ process is that the parameters $\{\theta\}$ are identifiable. An MA(q) process is $identifiable$ if Θ has unit root [10].

The $MA(q)$ model parameters $\{\phi\} = \{\theta_1.\theta_2, ..., \theta_q\}$ are estimated via the method of maximum likelihood and they are denoted as $\{\hat{\theta}\} = \{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_q\}$. The appropriate value of q is chosen via the AIC method. Once $\{\hat{\theta}\}$ are estimated and p is selected, the fitted values, denoted by $\hat{y}_{1:T}$, are estimated from the fitted model

$$\hat{y}_t = \sum_{j=1}^{q} \hat{\theta}_j \hat{\varepsilon}_{j-q} \tag{4.7}$$

where

$\hat{\varepsilon}_t = y_t - \hat{y}_{j-q}$ for $t = p+1, ..., T$

$\hat{\varepsilon}_1, \hat{\varepsilon}_2, ..., \hat{\varepsilon}_p \overset{iid}{\sim} normal(0, v)$ [10] .

Forecast Function

Suppose we fit $y_{1:365}$ to an $AR(p)$ model. Once the value of p is selected and the model parameters are estimated, the fitted values are calculated directly from equation 4.4. The fitted values, denoted by $\hat{y}_{1:365}$, represent the estimated daily average transactional response

times from Day 1 to Day 365. Suppose we also fit $y_{1:365}$ to an $MA(q)$ model. The set of fitted values would be directly calculated from equation 4.7 and they would also represent the estimated daily average transactional response time for the 365 days.

It is of interest to forecast the daily average response times for the next l days ($t = 366, ..., 365 + l$). The set of forecasted values are denoted by $\hat{y}_{366:365+l} = \{\hat{y}_{366}, ..., \hat{y}_{365+l}\}$. Each of the forecasted values are calculated directly from the forecast function. The forecast function is a linear combination of the estimated model parameters and the corresponding fitted values. Whether $\{\hat{\phi}\} = \{\hat{\phi}_1, \hat{\phi}_2, ..., \hat{\phi}_p\}$ or $\{\hat{\theta}\} = \{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_q\}$ are selected as the forecast function's model parameters depends on whether the fitted values are a result of an $AR(p)$ or $MA(q)$ fit [10].

The function in R that is used to forecast future values is $forecast()$ and it is from the forecast package in R [6]. This function is based off of the algorithm developed by Peiris and Perera (1988) [9].

Mixed Models

Autoregressive Moving Average Model: $ARMA(p, q)$

In time series analysis, there exists mixed models used to explain the dependency relationships in time series process. An autoregressive moving average model with orders p and q ($ARMA(p, q)$) is a mixed model that combines an autoregressive model with a moving average model. The parameters p and q represent the orders of the autoregressive and moving average parts, respectively. Causality and identifiability are important assumptions for the parameters in an $ARMA(p, q)$ model [10].

The time series process $\boldsymbol{Y_{1:T}}$, is an $ARMA(p, q)$ process if each $\boldsymbol{Y_t}$ arises from the autoregressive moving average time series model [10]

$$\boldsymbol{Y_t} = \sum_{j=1}^{p} \phi_j \boldsymbol{Y_{t-j}} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t, \tag{4.8}$$

where

ϕ_j = model parameter for the autoregressive term $\boldsymbol{Y_{t-j}}$

θ_j = model parameter for the moving average term ε_{t-j}

ε_t = stationary error at time t.

The autoregressive moving average time series model can also be written in terms of the characteristic equations 4.2 and 4.6 and the back shift operator B (eqn. 2.13) [10]

$$\Phi(B)\boldsymbol{Y_t} = \Theta(B)\varepsilon_t \tag{4.9}$$

where

$$\Phi(B) = (1 - \phi_1 B^1 - ... - \phi_p B^p) \text{ and } \Theta(z) = (1 - \theta_1 B^1 - ... - \theta_p B^q). \tag{4.10}$$

Autoregressive Integrated Moving Average Model: $ARIMA$

As mentioned before, causality and identifiability are important assumptions when fitting a time series to either an $AR(p)$, $MA(q)$ or $ARMA(p, q)$ model. If a time series is non-stationary, these assumptions are rarely met. Since it is frequent that time series processes are non-stationary with trends and seasonal patterns, there exists an autoregressive integrated moving average model ($ARIMA$) that is used to analyze the dependency within a non-stationary time series process. There exists two types of $ARIMA$ models; a non-seasonal autoregressive integrated moving average model with orders p, q and d ($ARIMA(p, q, d)$) and a seasonal autoregressive integrated moving average model with orders p, q, d, P, Q, D, and s ($ARIMA(p, q, d)(P, Q, D)_{[s]}$) [7].

Non-Seasonal $ARIMA(p, q, d)$ Model

If the observed daily average response times from $y_{1:365}$ increased over the 365 day period, we would assume that $y_{1:365}$ is a realization of a time series process that is non-stationary with increasing trend. The non-seasonal $ARIMA(p, q, d)$ model is appropriate for modeling a time series that appears to be non-stationary with trend. The $ARIMA(p, q, d)$ differs from the $ARMA(p, q)$ in that the prior applies the non-seasonal differencing operator to the time series in order to remove non-stationary trend [7].

The time series process $\boldsymbol{Y_{1:t}}$ is a non-seasonal $ARIMA(p, q, d)$ process if each $\boldsymbol{Y_t}$ arises from the autoregressive integrated moving average time series model [7]

$$\Phi(B)D^d\boldsymbol{Y_t} = \Theta(B)\varepsilon_t \tag{4.11}$$

where

ε_t = stationary error at time t.

D^d = non-seasonal differencing technique defined in equation 2.12.

It is assumed that the ε_t's are independent and identically distributed with mean zero and variance equal to v. The functions Φ and Θ are the characteristic function of the autoregressive and moving average parts defined in equation 4.10 [7].

Seasonal $ARIMA(p,q,d)[P,D,Q]_{[s]}$ Model

If the values of observed daily average response times from $y_{1:365}$ depended on the month they were observed, we would assume that $y_{1:365}$ is a realization of a time series process that is non-stationary with trend and seasonal pattern. The seasonal $ARIMA(p,q,d)(P,D,Q)_{[s]}$ model is appropriate for modeling a time series that appears to be non-stationary with trend and seasonal pattern with period s. The $ARIMA(p,q,d)(P,D,Q)_{[s]}$ applies a combination of the non-seasonal and seasonal differencing operators to the time series in order to remove the non-stationarity and explain the dependencies that exits due to the seasonal patterns [7].

The time series process $\boldsymbol{Y_{1:T}}$ is a seasonal $ARIMA(p,q,d)(P,D,Q)_{[s]}$ process if each $\boldsymbol{Y_t}$ arises from the autoregressive integrated moving average time series model [7]

$$\widetilde{\Phi}(B^s)\Phi(B)D_s^D D^d \boldsymbol{Y_t} = \widetilde{\Theta}(B^s)\Theta(B)\varepsilon_t, \tag{4.12}$$

where

ε_t = stationary error at time t.

D_d^s = seasonal differencing operator defined in equation 2.15.

It is assumed that the ε_t's are independent and identically distributed with mean zero and variance equal to v. The functions $\Phi(B)$ and $\Theta(B)$ are defined in equation 4.10. The functions $\widetilde{\Phi}$ and $\widetilde{\Theta}$ are the seasonal characteristic functions for the autoregressive and moving parts with degree P and Q, respectively, and they are defined by [7]

$$\Phi(B^s) = [1 - \widetilde{\Phi}_1(B^s)^1 - ... - \widetilde{\Phi}_p(B^s)^P] \text{ and } \Theta(z) = [1 - \widetilde{\Theta}_1(B^s)^1 - ... - \widetilde{\Theta}_p(B^s)^Q]. \tag{4.13}$$

Model Selection

Saturated Model.

Let M denote a set of models that are being analyzed in a time series model selection. Determining which model in M is the most appropriate to model the dependency relationships in a time series is a difficult task. There exists a hierarchy of models in M and the saturated model is the "largest model" because it can explain all other models in M. All other models in M can be written in terms of the saturated model because they are reduced versions of the saturated model. Before selecting a model, it is important to test all reduced models in M against the saturated model in M.

Let M be the set of the time series models previously discussed

$$M = \{AR(p),\ MA(q),\ ARMA(p,q),\ ARIMA(p,d,q)\},\ ARIMA(p,q,d)(P,D,Q)[s]\}.$$

The saturated model in M is the seasonal $ARIMA(p,q,d)(P,D,Q)[s]$ model because every reduced model in M can be written in terms of the seasonal ARIMA model (Eq. 4.12). For example, suppose $\boldsymbol{Y_t}$ arises from the seasonal ARIMA process with $P = Q = D = 0$ and $s = 1$, $ARIMA(p,q,d)(0,0,0)[1]$. $\boldsymbol{Y_t}$ is written as

$$\widetilde{\Phi}(B^1)\Phi(B)D_1^0 D^d \boldsymbol{Y_t} = \widetilde{\Theta}(B^1)\Theta(B)\varepsilon_t$$

$$\Rightarrow \widetilde{\Phi}(B^1)\Phi(B)(1-B^1)^0 D^d \boldsymbol{Y_t} = \widetilde{\Theta}(B^1)\Theta(B)\varepsilon_t$$

$$\Rightarrow \widetilde{\Phi}(B^1)\Phi(B)D^d \boldsymbol{Y_t} = \widetilde{\Theta}(B^1)\Theta(B)\varepsilon_t.$$

Since $\widetilde{\Phi}$ and $\widetilde{\Theta}$ are polynomials of order $P = 0$ and $Q = 0$, $\widetilde{\Phi}(B^1) = \widetilde{\Theta}(B^1) = 1$. Therefore, $\boldsymbol{Y_t}$ arises from

$$\Phi(B)D^d \boldsymbol{Y_t} = \Theta(B)\varepsilon_t. \tag{4.14}$$

Since equation 4.14 is equivalent to equation 4.11, $\boldsymbol{Y_t}$ arises from the non-seasonal $ARIMA(p,d,q)$ process, implying that $ARIMA(p,q,d)(0,0,0)[1] = ARIMA(p,q,d)$.

In most cases, all reduced models in M can be written in terms of other reduced models. For example, $AR(p)$ can be written as $ARMA(p, 0)$. Table 4.1 shows how the models in M are related.

Table 4.1: Relationships between models in M

Reduced Models	$ARMA$	Non-Seasonal ARIMA	Seasonal ARIMA
$AR(p) =$	$ARMA(p, 0) =$	$ARIMA(p, 0, 0) =$	$ARIMA(p, 0, 0)(0, 0, 0)[1]$
$MA(q) =$	$ARMA(0, q) =$	$ARIMA(0, q, 0) =$	$ARIMA(0, q, 0)(0, 0, 0)[1]$
	$ARMA(p, q) =$	$ARIMA(p, q, 0) =$	$ARIMA(p, q, 0)(0, 0, 0)[1]$
		$ARIMA(p, q, d) =$	$ARIMA(p, q, d)(0, 0, 0)[1]$

Estimation and Forecasting of ARIMA Models

We have shown that every reduced model in M can be written in the terms of the saturated model. Therefore, the seasonal $ARIMA(p, q, d)(P, Q, D)[s]$ is the starting point for selecting a model that is used to analyze the dependency relationships in a time series and to forecast future values. In R, the function $auto.arima()$ uses an algorithm developed by $Hyndman$ and $Khandakar$ (2008) in order to select the parameters and orders of $ARIMA(p, q, d)(P, Q, D)[s]$ [7]. A short explanation of this algorithm can be found in Appendix A and a more in depth explanation of the algorithm can be found in $Hyndman$ and $Khandakar$ (2008) [7]. For this analysis, the technique that is used to estimate the parameters is called Conditional Sums of Squares, a technique developed by Box and Jenkins (1970). Again, the $forecast()$ is used to forecast future values.

Chapter 5

Implementation To Oracle eBusiness Suite Application

Introduction

A transactional test is performed every five minutes on the Oracle eBusiness Suite application in order to monitor the application's transactional response time. These tests measure the amount of time it takes the application to receive a request and return a result. These measurements are stored in Nagios. Nagios is an open source network monitoring software application. A script was created in order to extract the data for this analysis from Nagios.

The transactional response times that were recorded over a 386 day period are used for this analysis. Each recording contains two elements; an unformatted time stamp (GMT time zone) and its corresponding transactional response time (seconds). The recordings observed from Day 1 to Day 365 make up Data Set 1 and are used to estimate the parameters of the GEV distribution and ARIMA model. The recordings observed from Day 366 to Day 386 are used to test the accuracy of these estimations.

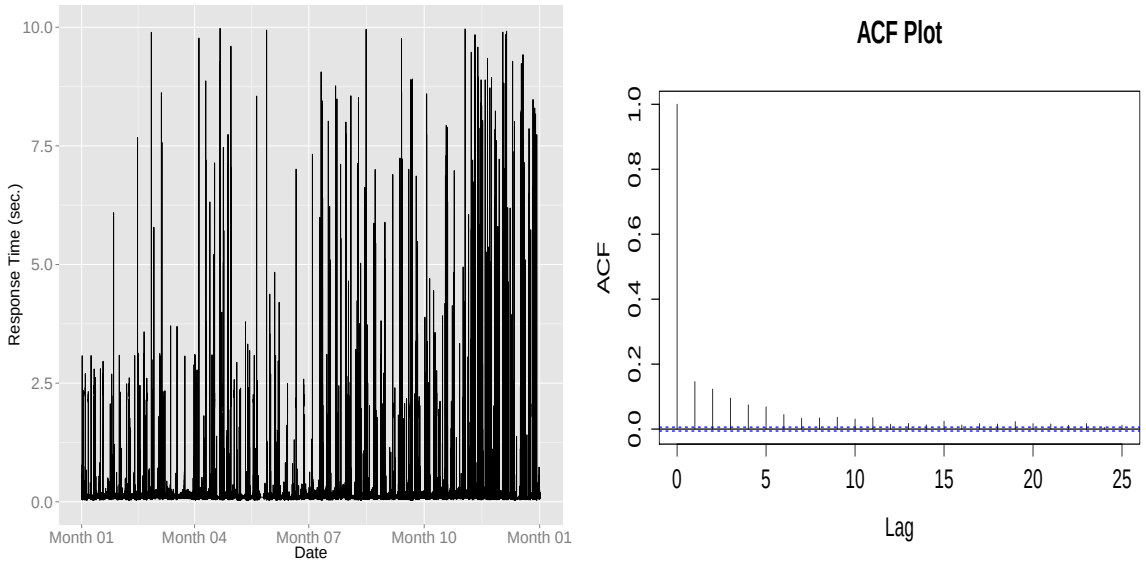
Daily Minimum Transactional Response Time

Checking for Stationarity

In order to determine whether a stationary GEV or a non-stationary GEV distribution will be used to estimate the distribution of the minimum transactional response time for the Oracle eBusiness Suite application, the assumption of stationary is investigated in Data Set 1.

Figure 5.1 contains the time series and ACF plots of the transactional response times observed from Day 1 to Day 365 (Data Set 1). In Figure 5.1(a), the response times appear

to increase from Month 10 - Month 12; this is evidence of a non-constant mean. In Figure 5.1(b), as lag increases, there continues to be ACF values that are greater than the level of significance (blue lines). Both of the plots in Figure 5.1 suggest that the time series is non-stationary. The results of the KPSS test return a p-value of .01. This indicates that there is enough evidence to reject the null hypothesis of unit root and assume that the data is non-stationary.



(a) Time Series Plot

(b) ACF Plot

Figure 5.1: Time series and ACF plots of every observed response times from Day 1 to Day 365 for the Oracle eBusiness suiteapplication (Data Set 1).

Block Minima Method

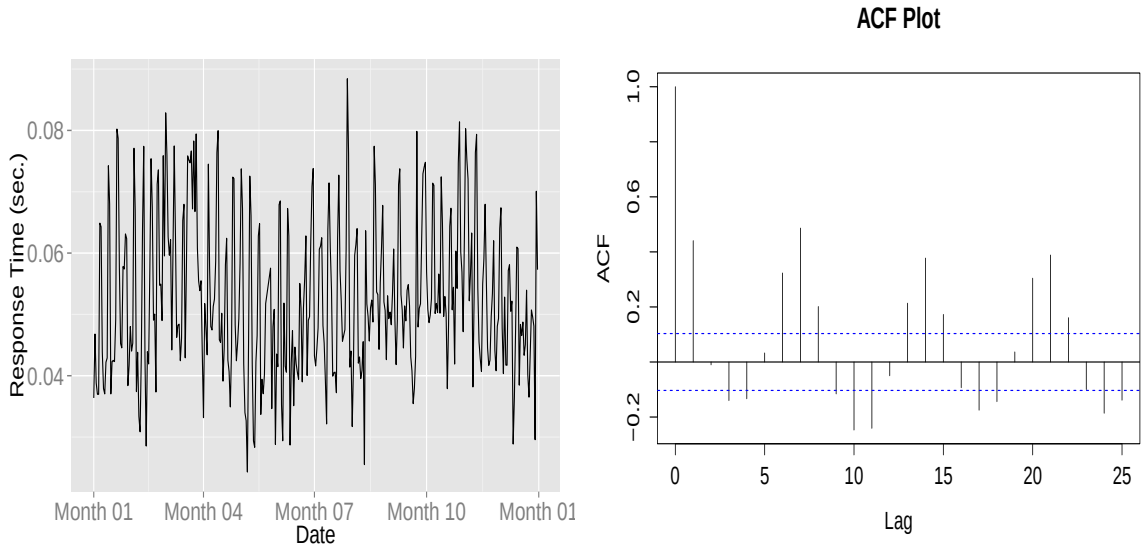
Data Set 1 is a time series of observations that are equally spaced by five minutes. The first 288 observations in Data Set 1 are the response times recorded during Day 1, the second 288 observations represent the response times recorded during Day 2,..., and the last 288 observations are the response times recorded during Day 365. Since the goal of this analysis is to estimate the distribution of the *daily* minimum response time, the Block Minima Method will be implemented by partitioning Data Set 1 into 365 non-overlapping blocks, where each blocks contains 288 observations.

The minimum values from each of the 365 blocks make up Data Set 2. The function in R that is used to perform this Block Minima Method is *ddply*() [13]. Data Set 2 represents the daily minimum transactional response times from Day 1 to Day 365 and is used to estimate the time dependent parameters in the non-stationary GEV distribution.

Selection of Time Dependent Parameters

Since there is evidence that Data Set 1 is non-stationary, the non-stationary $GEV(\mu(t,s), \sigma(t,s), \xi(t,s))$ distribution with non constant parameters will be used to estimate the distribution for the daily minimum transactional response time for the Oracle eBusiness Suite application. The non-stationary $GEV(\mu(t,s), \sigma(t,s), \xi(t,s))$ distribution and the selection of its time-dependent parameters are explained in detail in Section 3 and all possible time dependent parameters are listed in Table 3.1.

For this analysis, the selection of the parameter functions depends on the type of non-stationarity that appears in the time series and ACF plots of Data Set 2. If the time series plot reveals non-stationarity with trend, all of the parameters will be linear functions that are dependent on time (t). If the plots reveals non-stationarity with seasonal pattern, all of the parameters will be sinusoidal functions that are dependent on seasonal period (s). If the time series plot reveals non-stationarity with trend and seasonal pattern, all of the parameters will be a functions that are a sum of a linear and sinusoidal functions that are dependent on t and s.



(a) Time Series Plot

(b) ACF Plot

Figure 5.2: Time series and ACF plots of the observed daily minimum response times from Day 1 to Day 365 for the Oracle eBusiness suiteapplication (Data Set 2).

Figure 5.2 contains the time series and ACF plots of the time series from Data Set 2. The seasonal pattern in both Figure 5.2(a) and 5.2(b) suggest that there is a monthly seasonal pattern in Data Set 2. Figure 5.2(a) also suggest that the daily minimum response time decreases over time. Since these plots suggest that the time series in Data Set 2 is non-

stationary with decreasing trend and monthly seasonal pattern, the non-constant parameters in the non-stationary $GEV(\mu(t,s), \sigma(t,s), \xi(t,s))$ distribution will be the following functions

$$\mu(t, c_s) = \mu_o + \mu_1 t + \mu_{sin} sin(\omega c_s) + \mu_{cos} cos(\omega c_s) \tag{5.1}$$

$$\sigma(t, c_s) = \sigma_o + \sigma_1 t + \sigma_{sin} sin(\omega c_s) + \sigma_{cos} cos(\omega c_s) \tag{5.2}$$

$$\xi(t, c_s) = \xi_o + \xi_1 t + \xi_{sin} sin(\omega c_s) + \xi_{cos} cos(\omega c_s) \tag{5.3}$$

where $\theta_\mu = \{\mu_o, \mu_1, \mu_{sin}, \mu_{cos}\}$, $\theta_\sigma = \{\sigma_o, \sigma_1, \sigma_{sin}, \sigma_{cos}\}$, and $\theta_\xi = \{\xi_o, \xi_1, \xi_{sin}, \xi_{cos}\}$ are the sets of intercepts and coefficients for the functions $\mu(t, c_s), \sigma(t, c_s),$ and $\xi(t, c_s)$, respectively and $\omega = \frac{2\pi}{365.25}$. The variable c_s represents the monthly seasonal pattern in the time series and denotes the center of the s-th period counted in days starting from the beginning of the year. Since the plots in Figure 5.2 suggest that there is a monthly period, s =12. The center day of month s is denoted by c_s. For example, the center day of January is January 15th. Since, January 15th is the 15th day of the year, $c_1 = 15$. Table 5.1 defines all the center days of the months.

Table 5.1: Description of the center days of the months.

c_s	Center Day of Month s	Day of the Year
c_1	January 15	15
c_2	February 14	45
c_3	March 15	74
c_4	April 15	105
c_5	May 15	135
c_6	June 14	166
c_7	July 15	196
c_8	August 15	227
c_9	September 15	258
c_{10}	October 14	288
c_{11}	November 15	319
c_{12}	December 15	349

Estimating the non-stationary GEV Distribution

Equation 3.18 is the transformation that takes the normalized reciprocal of the block minimum values. This transformation is applied to Data Set 2 and the data set is fitted to

the non-stationary GEV distribution in R via the function $gev.fit()$. Appropriate residual diagnostics were performed and the data appears to fit the distribution appropriately.

Table 5.2 contains the estimated intercepts and parameters coefficients. With these estimated intercepts and coefficients, the estimated probability distribution for the daily minimum transactional response time is constructed. The estimated expected daily minimum response times from Day 1 to Day 365 (Eq. 3.19), the corresponding 95% lower confidence limit (Eq. 3.20) and the median daily minimum response time (Eq. 3.21) are estimated from the estimated parameters in Table 5.2.

Table 5.2: The estimated intercepts and parameter coefficients for the non-stationary GEV distribution.

Parameter	Set of Parameters	Intercept	Time	Sin	Cosine
Location: μ	$\hat{\theta}_\mu$	$\hat{\mu}_o = -0.3782$	$\hat{\mu}_1 = 0.0685$	$\hat{\mu}_{sin} = -0.0775$	$\hat{\mu}_{cos} = -0.0004$
Scale: σ	$\hat{\theta}_\sigma$	$\hat{\sigma}_o = 0.7987$	$\hat{\sigma}_1 = 0.0282$	$\hat{\sigma}_2 = 0.0847$	$\hat{\sigma}_3 = 0.0001$
Shape: θ	$\hat{\theta}_\xi$	$\hat{\xi}_o = 0.0271$	$\hat{\xi}_1 = 0.0458$	$\hat{\xi}_2 = 0.1088$	$\hat{\xi}_3 = -0.0002$

Figure 5.3 is a time series plot of Data Set 2 (black dots). The estimated expected values, median and lower limit for the 95% confidence interval are represented by the red, blue and green lines, respectively. Since the parameters are dependent on both t and s, the values for these estimates are non constant. It appears that only a few of the observed daily minimum response time from Data Set 2 fall below the lower 95% limit.

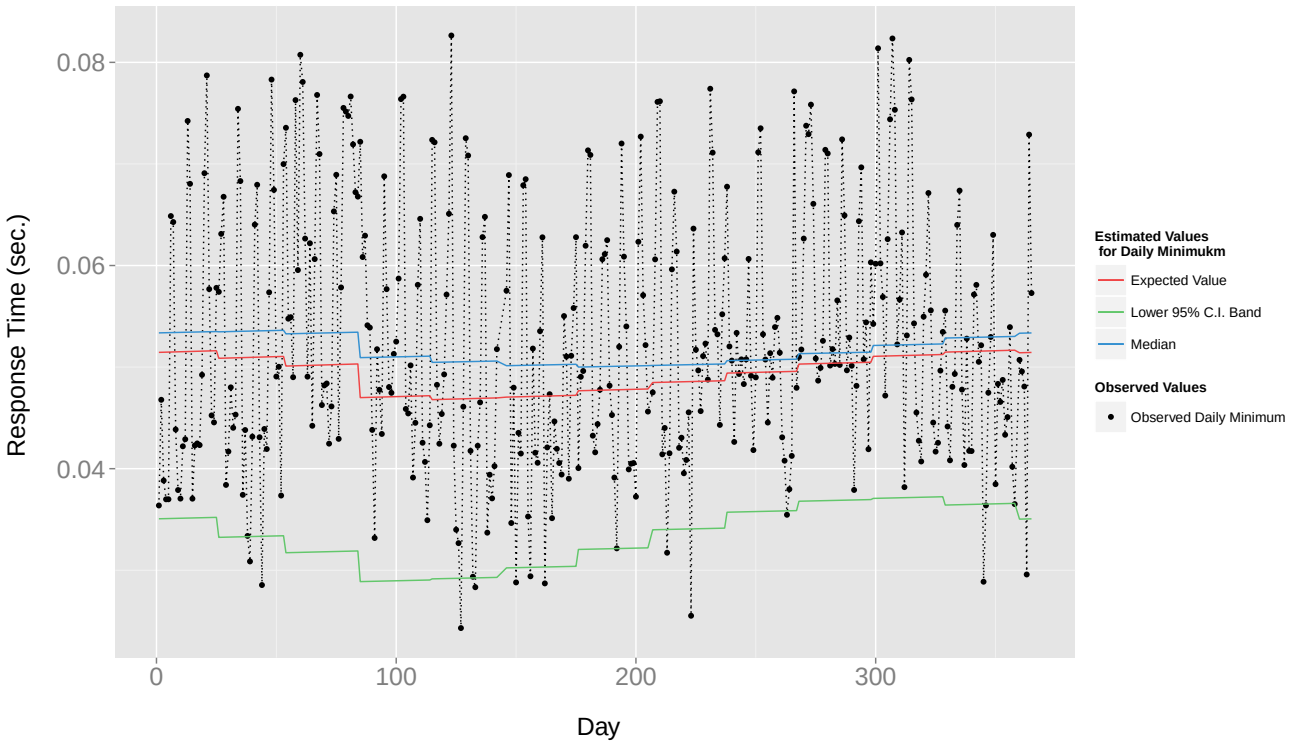


Figure 5.3: Time series plot of the observed daily response times from Day 1 to Day 365 for the Oracle eBusiness suiteapplication (Data Set 2). The time series plot contains the estimated values for the daily minimum response times. These estimated values are calculated from the estimated parameters in Table 5.2.

Alert System

One of the goals of fitting Data Set 2 to a non-stationary GEV distribution is to estimate future daily minimum response times and their corresponding statistics. These estimates can then be used to create an alert system that notifies when an incomplete transaction occurs in Oracle. Since Data Set 2 contains the daily minimum response times from Day 1 to Day 365, the daily minimum expected value, median and the lower 95% CI limit are estimated for Day 366 to Day 386. These estimates are calculated from the estimated parameters in Table 5.2. The observed response times from Day 366 to Day 386 are plotted against these estimates in order to demonstrate how an alert system can be created to detect anomalies that result in unsuccessful transactions.

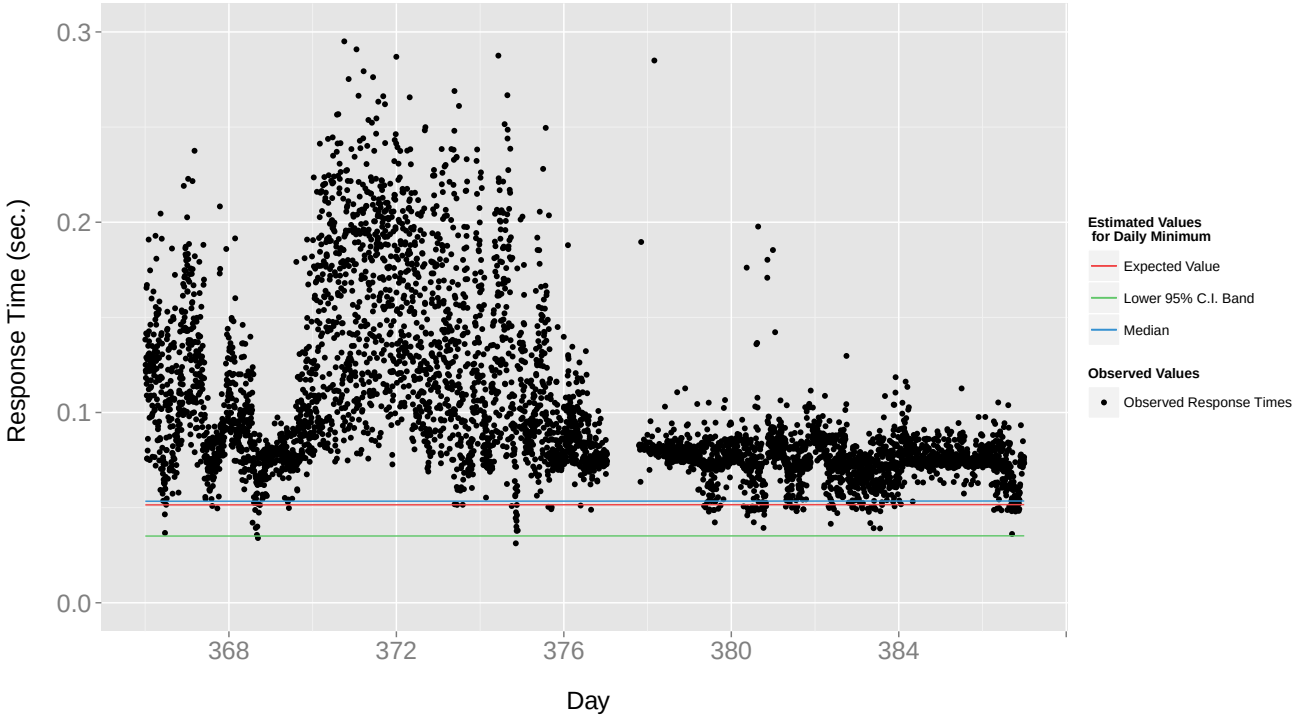


Figure 5.4: Time series plot of every observed response time from Day 366 to Day 386 for the Oracle eBusiness suite application. The time series plot contains the estimated values for the daily minimum response times. These estimated values are calculated from the estimated parameters in Table 5.2.

Figure 5.4 contains the estimated expected values (red line), median (blue line) and lower 95% CI limit (green line) for the daily minimum response times from Day 366 to Day 386. The lower 95% CI limit is interpreted as the shortest amount of time it would take for an Oracle eBusiness Suite application to successfully complete a transaction. Any response time that falls below this line suggest an unsuccessful transaction that occurred as a result of an

anomaly. The black dots represent every observed response time from Day 366 to Day 386. There are only three observed response times that fall below the lower limit. If this lower limit served as the lower threshold for an alert system,an alert would have been sent out the moment these response times occurred. This alert would have allowed for quicker anomaly detection.

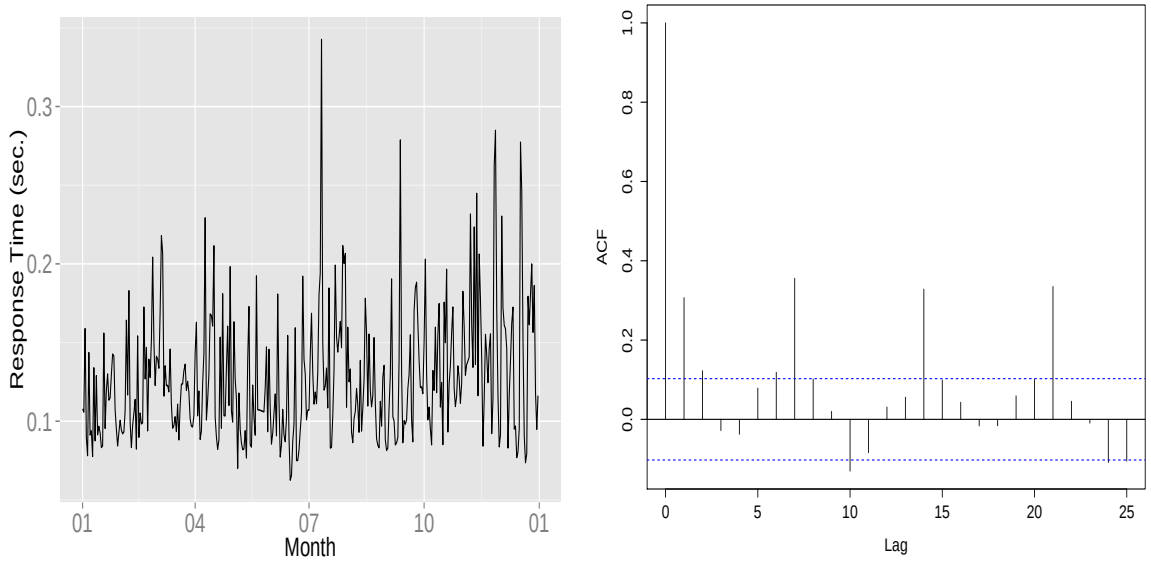
We have show the importance of being alerted when a response is shorter than usual as a result of an incomplete transaction. It is also important to be alerted when response times are longer than usual. Therefore, the daily average response time will be modeled in order to estimate an upper threshold for Oracle transactional response times.

Daily Average Response Time

Data Set 3 contains the daily average transactional response time from Day 1 to Day 365. It was created in R by applying the function *ddply*() to Data Set 1. Data Set 3 is used to estimate and forecast the daily average response time for Oracle. The assumption of stationarity is investigated in order to determine the most appropriate time series model for Data Set 3.

Checking for Stationarity

Figure 5.5 contains the time series and ACF plots of the daily average transactional response time for the Oracle eBusiness Suite application from Day 1 to Day 365 (Data Set 3). There appears to be a sinusoidal patter in Figure 5.5(a) and 5.5(b). This suggest that the time series is non-stationary with seasonal pattern. It also appears that the response times from Month 10 - Month 12 in Figure 5.5(a) have an increasing trend. The KPSS test for unit root was performed to algebraically check the stationary assumption. The results of the KPSS test return a p-value of .01. This indicates that there is enough evidence to reject the null hypothesis of unit root and assume that the data is non-stationary.



(a) Time Series Plot

(b) ACF Plot

Figure 5.5: Time series and ACF plots of the observed daily average response times from Day 1 to Day 365 for the Oracle eBusiness Suite application (Data Set 3)

Model Selection

Figure 5.5 and the results of the KPSS test indicate that the data is non-stationary with both increasing trend and seasonal pattern. Therefore, we will begin our model selection with the seasonal $ARIMA(p,q,d)[P,Q,D]_{[s]}$. Since our data is collected daily, we can investigate a yearly period ($s = 365$), a monthly period, ($s = 30$) and a weekly period ($s = 7$). However, since the data set contains exactly 365 observations, a model with $s = 365$ will return the same results as a model with $s = 1$ because there is only one observation per day of the year. Therefore, we will begin our model selection by comparing $ARIMA(p,q,d)[P,Q,D]_{[s]}$ models with $s = 30$ and $s = 7$.

The *auto.arima*() function in R is used to select the most appropriate time series model to describe the dependency relationship in the daily average transactional response times for the Oracle eBusiness Suite application. The results suggest that the model with the smallest AIC value is the seasonal autoregressive integrated moving average model $ARIMA(2,2,1)[2,1,0]_7$. This implies that the time series analyses arises from a seasonal ARIMA process where $\boldsymbol{Y_t}$ arises from equation 5.4. With some algebra, equation 5.5 can be reduced to equation 5.5.

$$\widetilde{\Phi}(B^7)\Phi(B)D_7^0 D^1 \boldsymbol{Y_t} = \widetilde{\Theta}(B^7)\Theta(B)\varepsilon_t \tag{5.4}$$

$$\Rightarrow (1-\widetilde{\phi}_1 B^7 - \widetilde{\phi}_2 B^{14})(1-\phi_1 B - \phi_2 B^2) D\boldsymbol{Y_t} = (1-\widetilde{\theta}_1 B^7)(1-\theta_1 B - \theta_2 B^2)\varepsilon_t$$

$$\Rightarrow (1 - \phi_1 B - \phi_2 B^2 - \widetilde{\phi}_1 B^7 + \phi_1\widetilde{\phi}_1 B^8 + \phi_2\widetilde{\phi}_1 B^9 - \widetilde{\phi}_2 B^{14} + \phi_1\widetilde{\phi}_2 B^{15} + \phi_2\widetilde{\phi}_2 B^{16})(\boldsymbol{Y_t} - \boldsymbol{Y_{t-1}}) =$$

$$(1 - \theta_1 B - \theta_2 B^2 - \widetilde{\theta}_1 B^7 + \widetilde{\theta}_1\theta_1 B^8 + \widetilde{\theta}_1\theta_2 B^9)\varepsilon_t$$

$$\Rightarrow \boldsymbol{Y_t} = (1+\phi_1)\boldsymbol{Y_{t-1}} + (\phi_2-\phi_1)\boldsymbol{Y_{t-2}} - \phi_2\boldsymbol{Y_{t-3}} + \widetilde{\phi}_1\boldsymbol{Y_{t-7}} - \widetilde{\phi}_1(1+\phi_1)\boldsymbol{Y_{t-8}} + \tag{5.5}$$

$$\widetilde{\phi}_1(-\phi_2+\phi_1)\boldsymbol{Y_{t-9}} + \phi_2\widetilde{\phi}_1\boldsymbol{Y_{t-10}} - \widetilde{\phi}_2\boldsymbol{Y_{t-14}} - \widetilde{\phi}_2(1+\phi_1)\boldsymbol{Y_{t-15}} + \widetilde{\phi}_2(\phi_1-\phi_2)\boldsymbol{Y_{t-16}} +$$

$$\phi_2\widetilde{\phi}_2\boldsymbol{Y_{t-17}} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \widetilde{\theta}_1\varepsilon_{t-7} + \widetilde{\theta}_1\theta_1\varepsilon_{t-8} + \widetilde{\theta}_1\theta_2\varepsilon_{t-9}$$

Dependency Relationships.

The selected orders for this model indicates that the random variable $\boldsymbol{Y_t}$ has a significant covariance with

$\{\boldsymbol{Y_{t-1}}, \boldsymbol{Y_{t-2}}, \boldsymbol{Y_{t-3}}, \boldsymbol{Y_{t-7}}, \boldsymbol{Y_{t-8}}, \boldsymbol{Y_{t-9}}, \boldsymbol{Y_{t-10}}, \boldsymbol{Y_{t-14}}, \boldsymbol{Y_{t-15}}, \boldsymbol{Y_{t-16}}, \boldsymbol{Y_{t-17}}\}$ and $\{\varepsilon_{t-1}, \varepsilon_{t-2}, \varepsilon_{t-7}, \varepsilon_{t-8}, \varepsilon_{t-9}\}$.

These significant covariance can be interpreted as follows. The selection of $p = 2$ implies that $\boldsymbol{Y_t}$ has a significant covariance with the 2 previous day's average response time ($\boldsymbol{Y_{t-1}}$ and $\boldsymbol{Y_{t-1}}$). The selection of $q = 2$ implies that $\boldsymbol{Y_t}$ has a significant covariance with the random errors from the two previous days (ε_{t-1} and ε_{t-2}). The seasonal order selection of $s = 7$ implies that $\boldsymbol{Y_t}$ has a significant covariances with random variables and random errors from previous weeks. For example, the selection of $P = 2$ implies that $\boldsymbol{Y_t}$ has a significant correlation with daily response times from the previous two weeks ($\boldsymbol{Y_{t-7}}$ and $\boldsymbol{Y_{t-14}}$). Since $p = 2$, $\boldsymbol{Y_t}$ also has a significant correlation with $\boldsymbol{Y_{t-8}}$, $\boldsymbol{Y_{t-9}}$, $\boldsymbol{Y_{t-15}}$ and $\boldsymbol{Y_{t-16}}$. Since $Q = 1$, $\boldsymbol{Y_t}$ has a significant correlation with the random error from the previous week (ε_{t-7}). Since $q = 2$, $\boldsymbol{Y_t}$ also has a significant correlation with ε_{t-8} and ε_{t-9}. Lastly, in order to remove the stationarity from the time series, an non-differencing operator with order $d = 1$ is applied to each of the random variables in the model. This adds the random variables $\boldsymbol{Y_{t-10}}$ and $\boldsymbol{Y_{t-17}}$ to the model.

Fitted Values

The estimated model parameters that are returned from fitting Data Set 3 to the seasonal $ARIMA(2,2,1)[2,1,0]_7$ model from equation 5.5 are in Table 5.3. Equation 5.6 represents the the fitted seasonal $ARIMA(2,2,1)[2,1,0]_7$ model.

Estimated Parameter	Value
$\hat{\phi}_1$	0.7417
$\hat{\phi}_2$	0.1986
$\hat{\widetilde{\phi}}_1$	0.9551
$\hat{\widetilde{\phi}}_2$	0.0359
$\hat{\theta}_1$	0.1063
$\hat{\theta}_1$	-0.7900
$\hat{\widetilde{\theta}}_1$	-0.8669

Table 5.3: The estimated model parameter coefficients for the seasonal $ARIMA(2,2,1)[2,1,0]_7$ model.

$$\Rightarrow \hat{y}_t \;=\; (1+\hat{\phi}_1)\hat{y}_{t-1} \;+\; (\hat{\phi}_2 - \hat{\phi}_1\hat{y}_{t-2} \;-\; \hat{\phi}_2\hat{y}_{t-3} \;+\; \hat{\tilde{\phi}}_1\hat{y}_{t-7} \;-\; \hat{\tilde{\phi}}_1(1+\hat{\phi}_1\hat{y}_{t-8} + \tag{5.6}$$

$$\hat{\tilde{\phi}}_1(-\hat{\phi}_2+\hat{\phi}_1)\hat{y}_{t-9} \;+\hat{\phi}_2\hat{\tilde{\phi}}_1\hat{y}_{t-10} - \;\hat{\tilde{\phi}}_2\hat{y}_{t-14} \;-\; \hat{\tilde{\phi}}_2(1+\hat{\phi}_1)\hat{y}_{t-15} \;+\; \hat{\tilde{\phi}}_2(\hat{\phi}_1-\hat{\phi}_2)\hat{y}_{t-16} +$$

$$\hat{\phi}_2\hat{\tilde{\phi}}_2\hat{y}_{t-17} + \;\hat{\varepsilon}_t \;-\; \hat{\theta}_1\hat{\varepsilon}_{t-1} \;-\; \hat{\theta}_2\hat{\varepsilon}_{t-2} \;-\; \hat{\tilde{\theta}}_1\hat{\varepsilon}_{t-7} \;+\; \hat{\tilde{\theta}}_1\hat{\theta}_1\hat{\varepsilon}_{t-8} \;+\; \hat{\tilde{\theta}}_1\hat{\theta}_2\hat{\varepsilon}_{t-9}$$

Once the estimated values from Table 5.3 are plugged into equation 5.6, the set of fitted value, denoted by $\hat{y}_{1:365}$, are calculated directly from equation 5.6. The set $\hat{y}_{1:365}$ represent the estimated daily average transactional response times from Day 1 to Day 365. Figure 5.6 is Data Set 3 (black dotted line) plotted against $\hat{y}_{1:365}$ (red line). The fitted values appear to model the data appropriately and residual diagnostics were analyzed and confirm an appropriate fit.

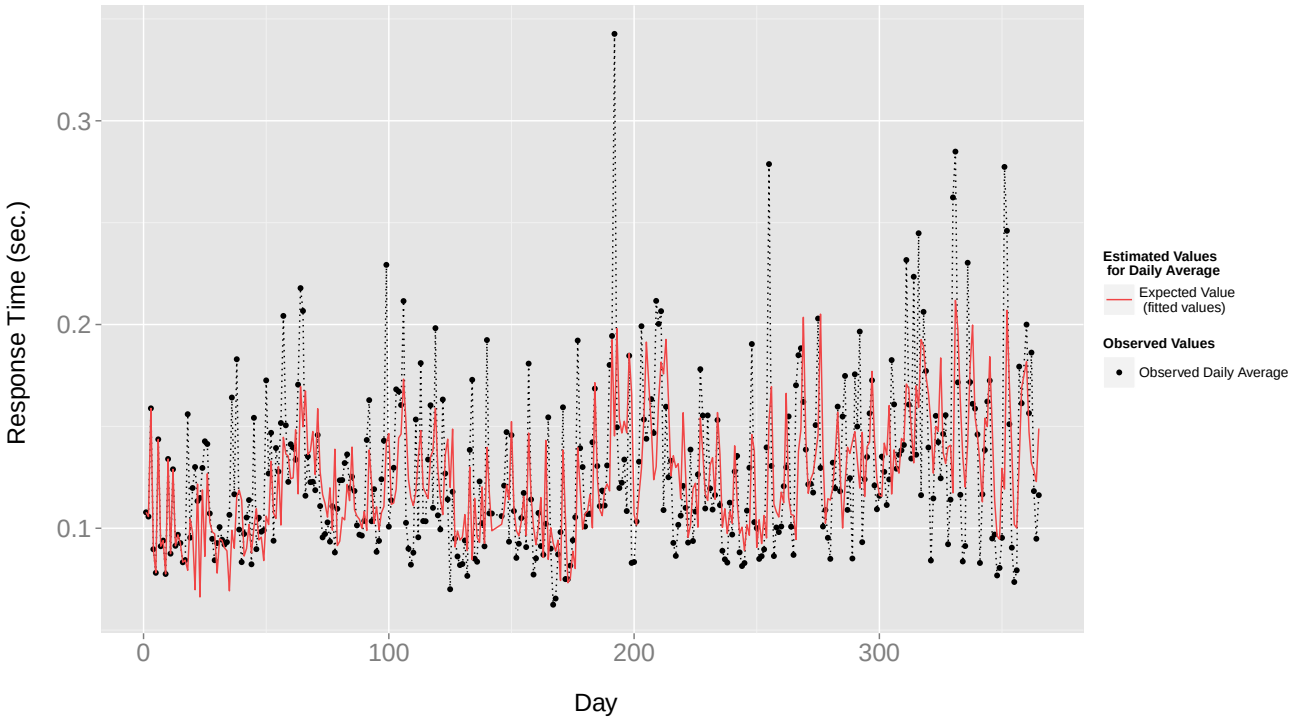


Figure 5.6: A time series plot of the observed daily average response times from Day 1 to Day 365 for the Oracle eBusiness Suite application (Data Set 3). The time series plot contains the estimated daily average values. These values are calculated from the fitted seasonal $ARIMA(2,2,1)[2,1,0]_7$ model (eqn. 5.6).

Alert System

Once the models parameters for the seasonal $ARIMA(1,2,1)[2,1,0]_{[7]}$ are estimated, the $forecast()$ function in R is used to forecast the daily average response times for Day 366 to Day 386 and their corresponding upper 95% confidence limits. The 95% confidence limit will be used as the upper threshold for an alert system that will be used to detect longer than usual response times.

Figure 5.7 contains estimated daily average response times for Day 366 to Day 386 (red line) and the corresponding upper 95% confidence limit (green line). The black dots represent the observed daily average response times from Day 366 to Day 386. Two of the observed daily average response times lie above the 95% upper confidence limit. If we let the 95% upper confidence limit be the upper threshold for an end of day alert system, then an alert would have been sent out notifying that there was a performance issue throughout these two days.

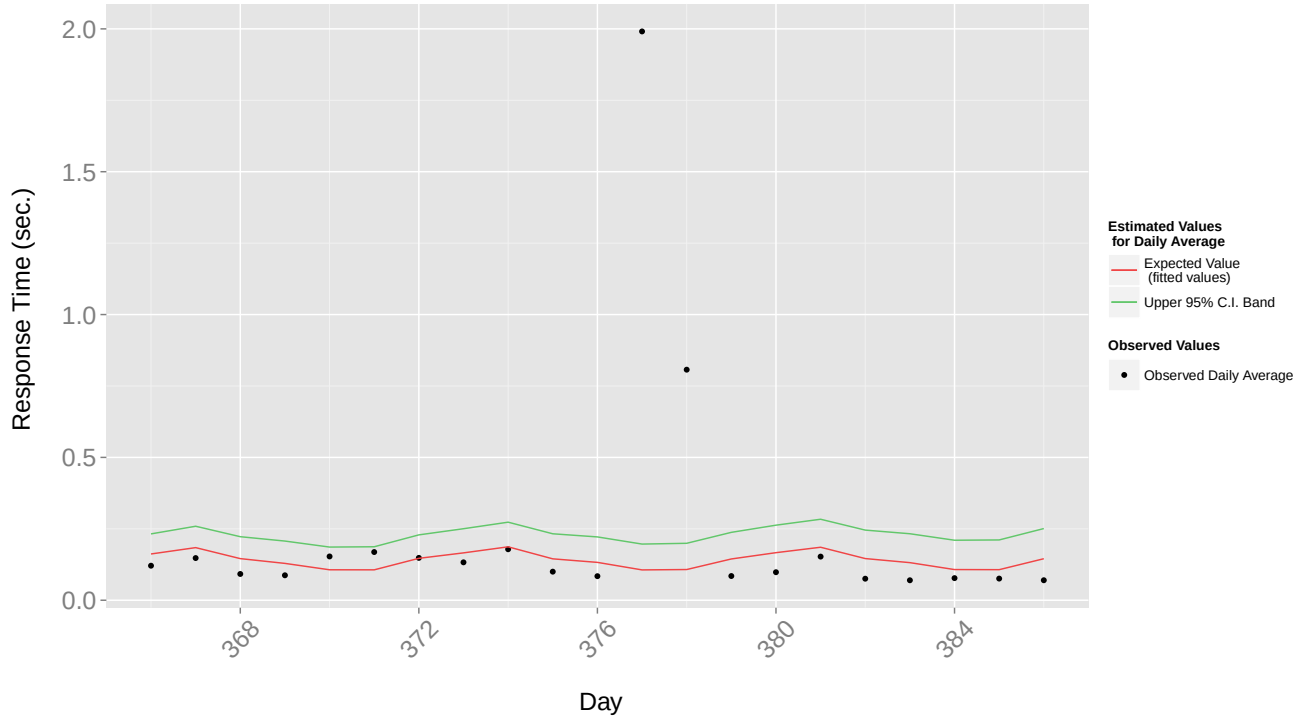


Figure 5.7: A time series plot of the observed daily average response times from Day 366 to Day 386 for the Oracle eBusiness Suite application. The time series plot contains the estimated daily average values. These values are calculated from the forecast function and the estimated model parameters from Table 5.3.

Using the upper 95% confidence limit as the upper threshold for an end of the day alert system is beneficial for solving performance issues. However, it is important to send out an alert the moment a pattern of long response times that are a result of a performance issue starts to occur. Therefore, the estimated hourly average response time will be estimated in order to have an hourly alert system.

Hourly Average Response Time

The methodology used to estimate the hourly average response time is identically to the methodology used to model the daily average response time. The data set that is used to estimate and forecast the daily average response time for the Oracle eBusiness Suite application is created in R by applying the function *ddply*() to Data Set 1. This returns a data set that contains the hourly average transactional response time from Day 1 to Day 365 (Data Set 4).

Checking for Stationarity

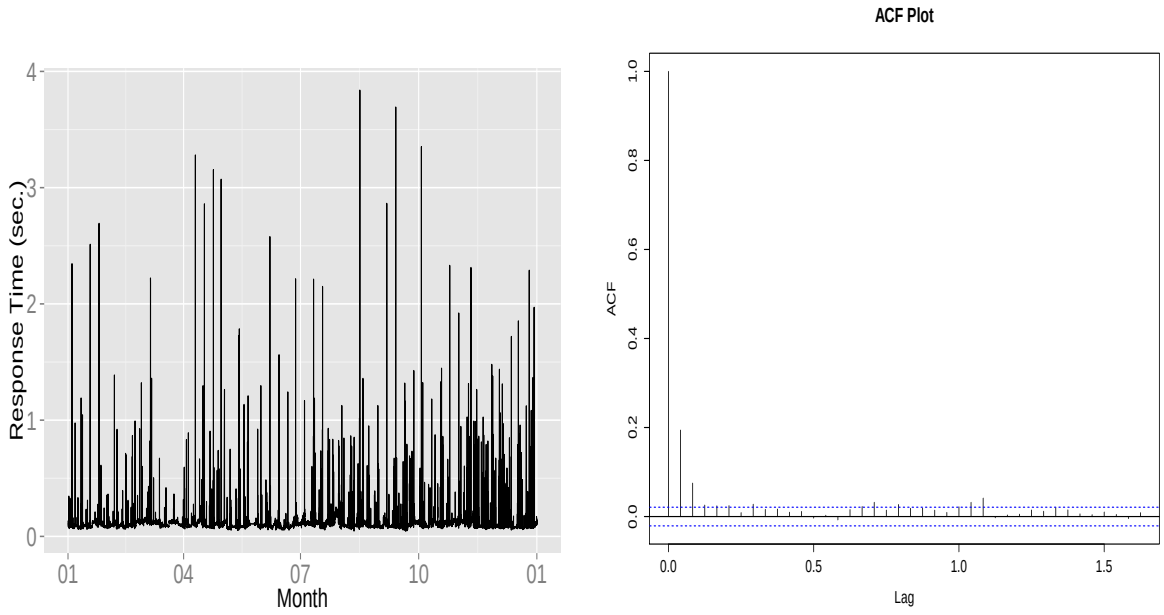
Figure 5.8 contains the time series and ACF plots of the hourly average transactional response time for the Oracle eBusiness Suite application from Day 1 to Day 365 (Data Set 4). Both Figures 5.8(a) and 5.8(b) appear to have a slight seasonal trend. The results of the KPSS test return a p-value of 0.01. This indicates that there is there is enough evidence to reject the null hypothesis of unit root and we may assume that the data is non-stationary. Since both the diagnostic plots and KPSS test suggest non-stationarity, we will begin our model selection with the $ARIMA(p,q,d)[P,Q,D]_{[s]}$ model.

Model Selection

The results of applying the *auto.arima*() function to Data Set 4 selects an $ARIMA(1,2,1)$ $(0,1,0)_{[24]}$ as the most appropriate time series model to describe the dependency relationship in the hourly average transactional response times for the Oracle eBusiness Suite application. This implies that $\boldsymbol{Y_t}$ arises from the autoregressive model

$$[1-\phi_1 B]D^1\boldsymbol{Y_t} = [1-\widetilde{\theta}_1 B^{24}][1-\theta_1 B-\theta_2 B^2]\varepsilon_t. \tag{5.7}$$

The selected model parameters imply that the differenced hourly average response time for the Oracle eBusiness Suite application depends only on the previous hour's average response time. Whereas, an hour's random error depends on the 2 previous hours' random errors and the random error from 24 hours prior.



(a) Time Series Plot

(b) ACF Plot

Figure 5.8: Time series and ACF plot of the observed hourly average response times from Day 1 to Day 365 for the Oracle eBusiness Suite application (Data Set 4).

Fitted Values

The fitted $ARIMA(1,2,1)(0,1,0)_{[24]}$ model that is used to calculate the estimated hourly average response times is

$$[1-\hat{\phi}_1 B]D^1\hat{y}_t = [1-\hat{\tilde{\theta}}_1 B^{24}][1-\hat{\theta}_1 B-\hat{\theta}_2 B^2]\hat{\varepsilon}_t, \tag{5.8}$$

where $\hat{\phi}_1 = 0.362$, $\hat{\theta}_1 = -1.1745$, $\hat{\theta}_2 = 0.1756$ and $\hat{\tilde{\theta}}_1 = 0.0119$ are the values of the estimated parameters. The value $\hat{y}_t$ represents the estimated average response time for hour t and $\hat{\varepsilon}_t$ represents the difference between the observed and estimated average response for hour t, where $\hat{\varepsilon}_t = y_t - \hat{y}_t$.

Figure 5.9 contains a plot of the estimated hourly average response times calculated from the fitted model (red line) against the observed hourly average response times (black dotted line) for the Oracle eBusiness Suite application from Day 1 to Day 365. The fitted values appear to fit the observed data well and appropriate residual diagnostics were performed and confirm an appropriate fit.

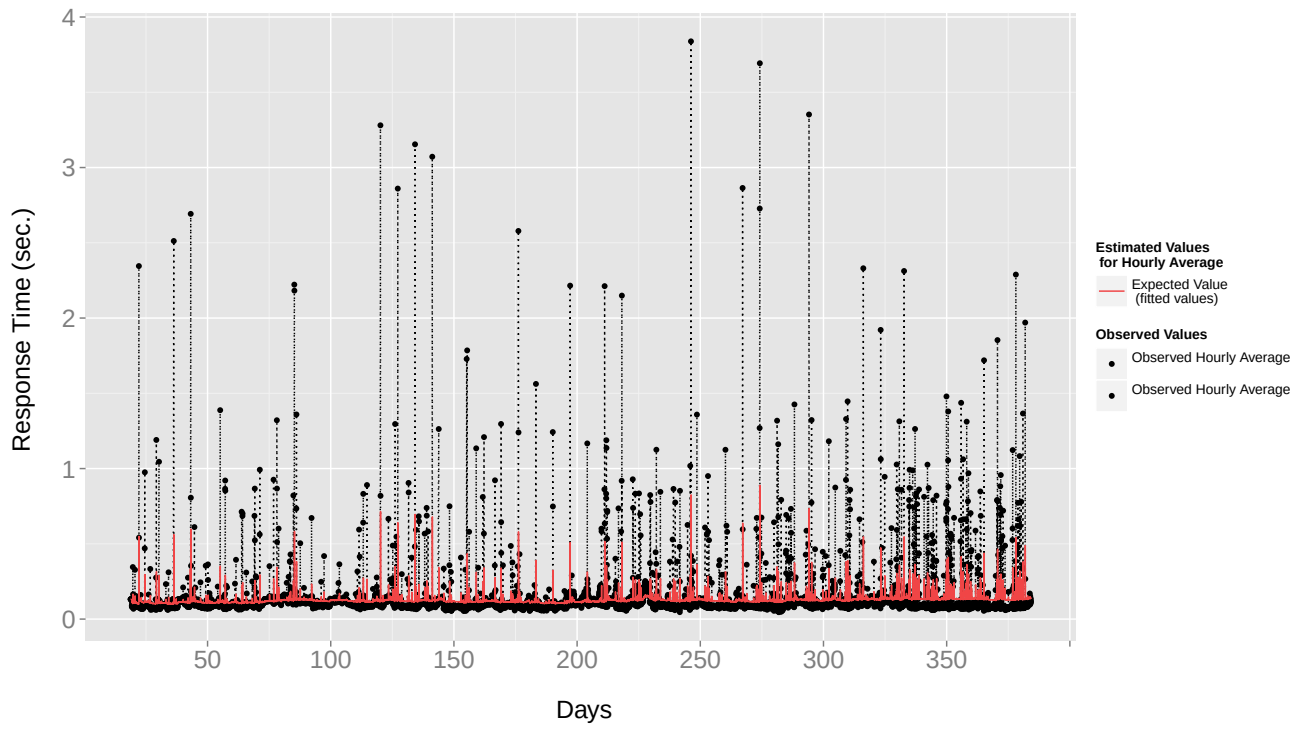


Figure 5.9: A time series plot of the observed hourly average response times from Day 1 to Day 365 for the Oracle eBusiness Suite application (Data Set 4). The time series plot contains the estimated hourly average values. These values are calculated from the fitted $ARIMA(1,2,1)(0,1,0)_{[24]}$ model (eqn. 5.8).

Alert System

Figure 5.10 contains a plot of the estimated hourly response times (red line) for Day 366 to Day 386 and the corresponding 95% upper limit (green line). The black dots represent the observed hourly average response times from Day 366 to Day 386. There appears to be a few hourly average response times that are greater than the 95% upper limit.

Most of the long hourly average response times were observed occurred between Days 378 and 379. The Oracle eBusiness Suite application was investigated after this analysis was performed and it was discovered that on Day 377, the application had a power outage that resulted in unrecorded response times and large response times from Days 378 and 379. If this alert system was implemented at the time of this power outage and the upper limit was used as the upper threshold from an alert system, an alarm would have been sent out the moment the first unusually long hourly average response time occurred.

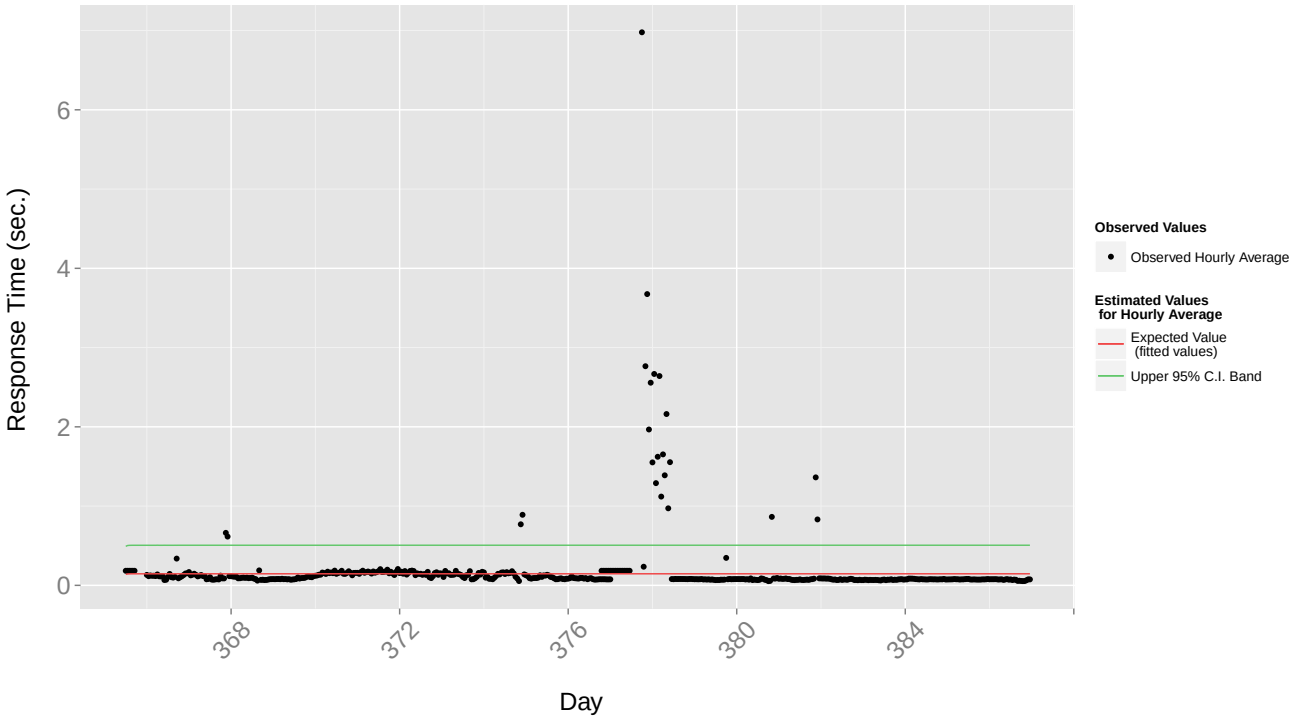


Figure 5.10: A time series plot of the observed hourly average response times from Day 366 to Day 386 for the Oracle eBusiness Suite application. The time series plot contains the estimated hourly average values. These values are calculated from the forecast function and the estimated model parameter from the fitted $ARIMA(1,2,1)(0,1,0)_{[24]}$ model (eqn. 5.8).

Chapter 6

Implementation to Weblogic c12 Server Application

Introduction

The data used to analyze Weblogic c12 Server application's transactional response time are from the transactional tests that are performed every five minutes on the Weblogic c12 Server application. Like the Oracle eBusiness suite application, the data is stored in Nagios. Data Set 5 contains the recorded transactional response times from the tests that were performed from Day 1 to Day 365. Each recording from this data set contains two elements, an unformatted time stamp (GMT time zone) and its corresponding transactional response time (seconds). The recorded data from Day 366 to 386 is used to test the accuracy of the estimations calculated from Data Set 5. Since the methodology used to perform this analysis is identical to the methodology described in Chapter 5, we will start by discussing the results of estimating the non-staionary GEV Distribution for the daily minimum response times.

Daily Minimum Response Time

Results

Figure 6.1 is a time series plot of the daily minimum response times from Day 1 to Day 365 for the Weblogic c12 Server application (black dotted line). It contains the estimated Weblogic c12 server application's expected daily minimum response times (red line), the corresponding 95% lower confidence limit (green line)and the estimated median (blue line). Appropriate residual diagnostics were performed and suggest that there is an appropriate fit. Figure 6.2 contains every observed response time from Day 366 to Day 386 and the forecasted daily minimum response time for the Weblogic c12 Server application. On Day 368, there appears to be six observed response times that fell below the lower limit. If the lower limit was used as the lower threshold for an alert system, an alarm would have been sent out the moment the first unusually short response time fell below the threshold. This would have allowed for a quicker and more accurate anomaly detection.

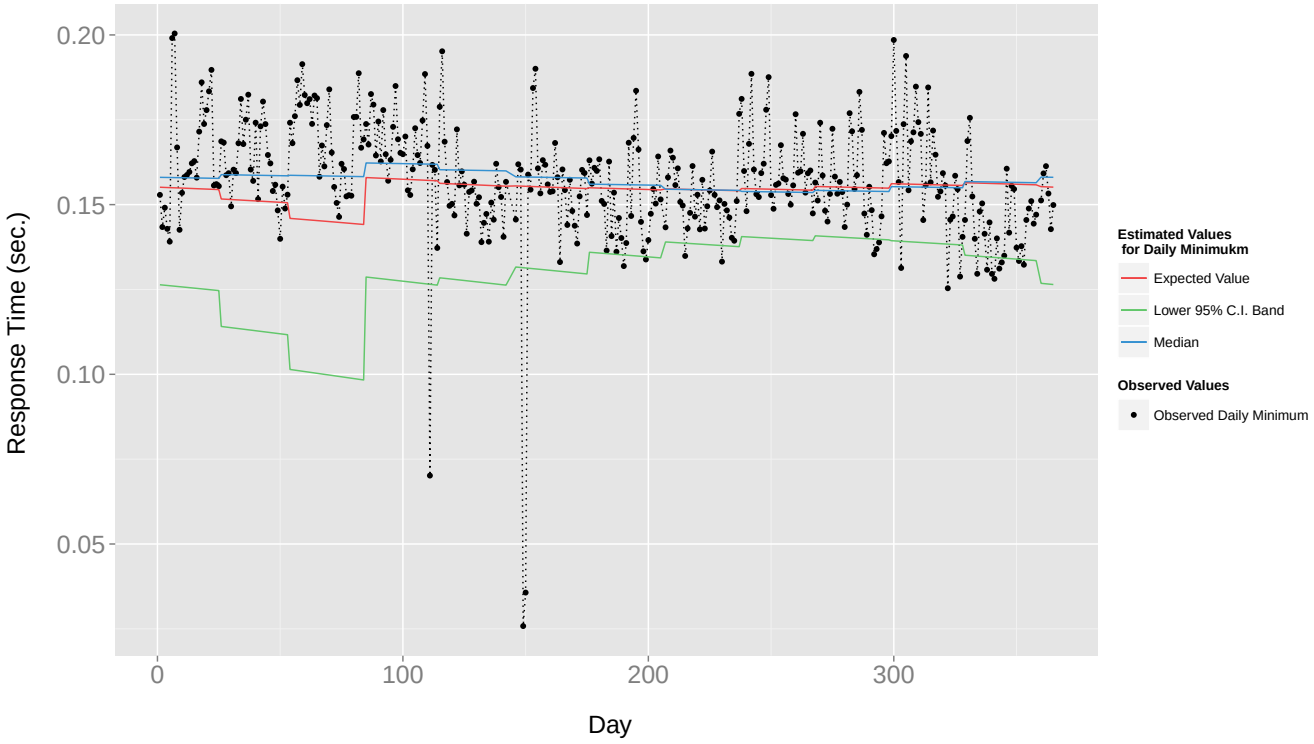


Figure 6.1: A time series plot of the observed daily minimum response times from Day 1 to Day 365 for the Weblogic c12 Server application. The time series plot contains the estimated daily minimum response times. These estimated values are calculated from the estimated parameters from the fitted non-stationary GEV distribution.

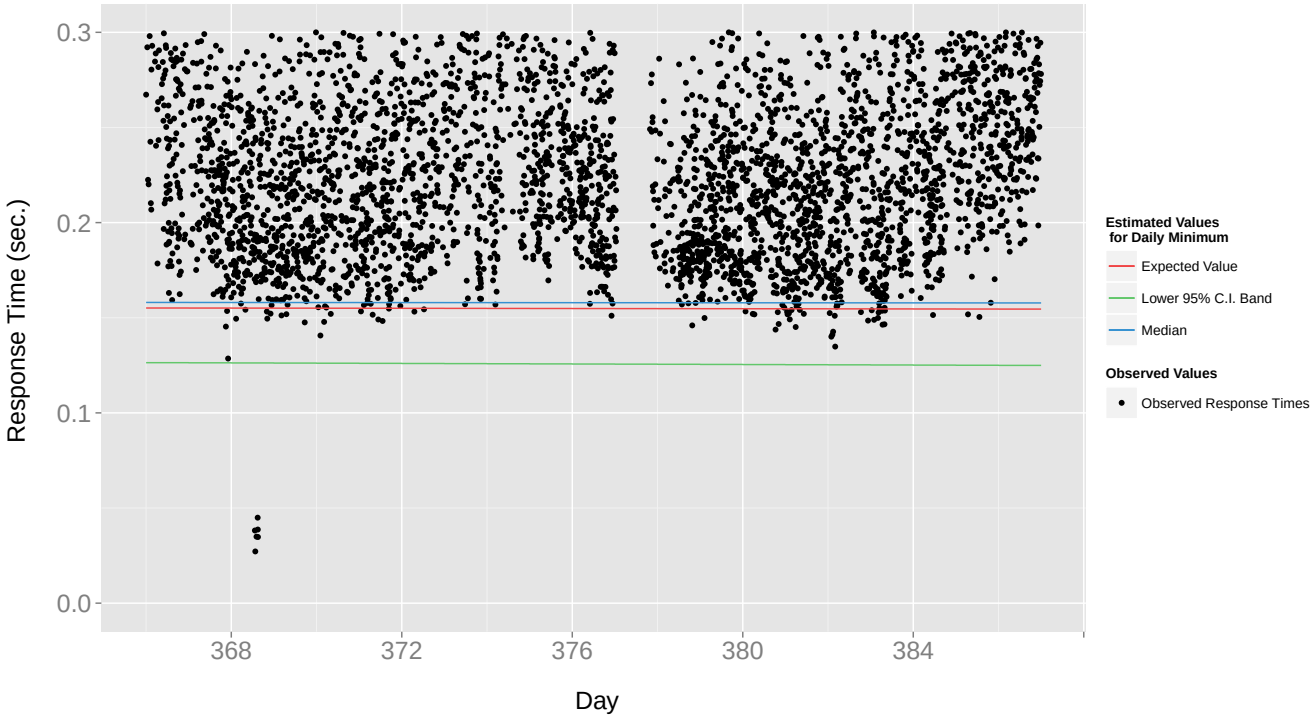


Figure 6.2: A time series plot of every observed response times from Day 366 to Day 386 for the Weblogic c12 Server application. The time series plot contains the estimated daily minimum response times. These estimated values are calculated from the estimated parameters from the fitted non-stationary GEV distribution.

Hourly Average Response Time

Results

Since having a continuous alert system is more beneficial than having an end of day alert system, we will only discuss the results of modeling Weblogic c12 server application's hourly average response time. The model that is selected as the most appropriate model to estimate the hourly average response times and forecast future response times is a non-seasonal $ARIMA(2, 2, 0)$ model.

Figure 6.3 is a time series plot of Weblogic c12 server application's observed hourly average response times for Day 1 to Day 365 (block dotted line). The red line represents the estimated expected values for the hourly average response times that were calculated from the fitted non-seasonal $ARIMA(2, 2, 0)$ model.

Figure 6.4 contains the foretasted hourly average response times for Day 366 to Day 386 and the corresponding 95% upper confidence limit. The black dots represent the observed hourly average response times from Day 366 to Day 386. There appears to be some hourly average response times that are greater than the upper limit; especially between Days 378 and 379. Days 378 and 379 are the dates that the power outage on Day 377 affected the Oracle eBussiness suite application's response times. Further investigation was done on the Weblogic c12 server application, and it was found that the power outage affected this application as well. If the upper limit served as the upper threshold for an alert system, a notification of these long response times would have been sent the moment they had started to occur; this would have allowed for quicker anomaly detection.

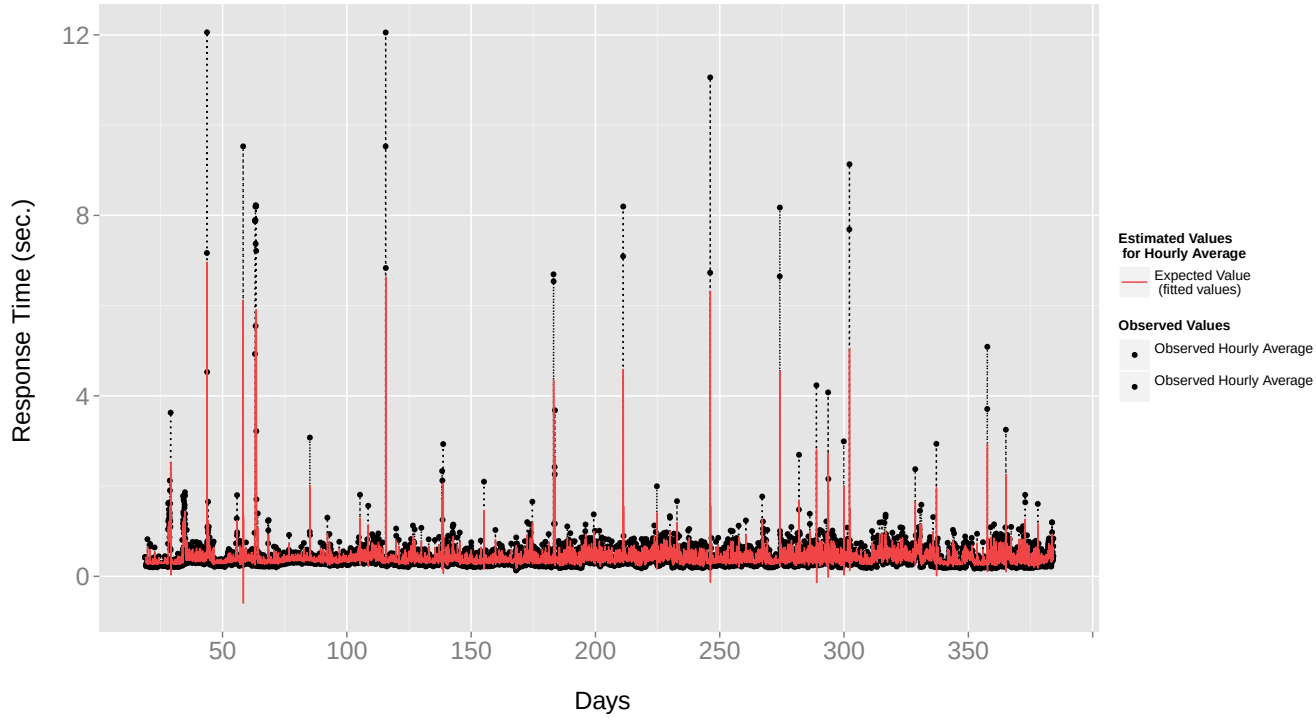


Figure 6.3: A time series plot of the observed hourly average response times from Day 1 to Day 365 for the Weblogic c12 Server application. The time series plot contains the estimated hourly average response times. These estimated values are calculated from the fitted non-seasonal $ARIMA(2, 2, 0)$ model.

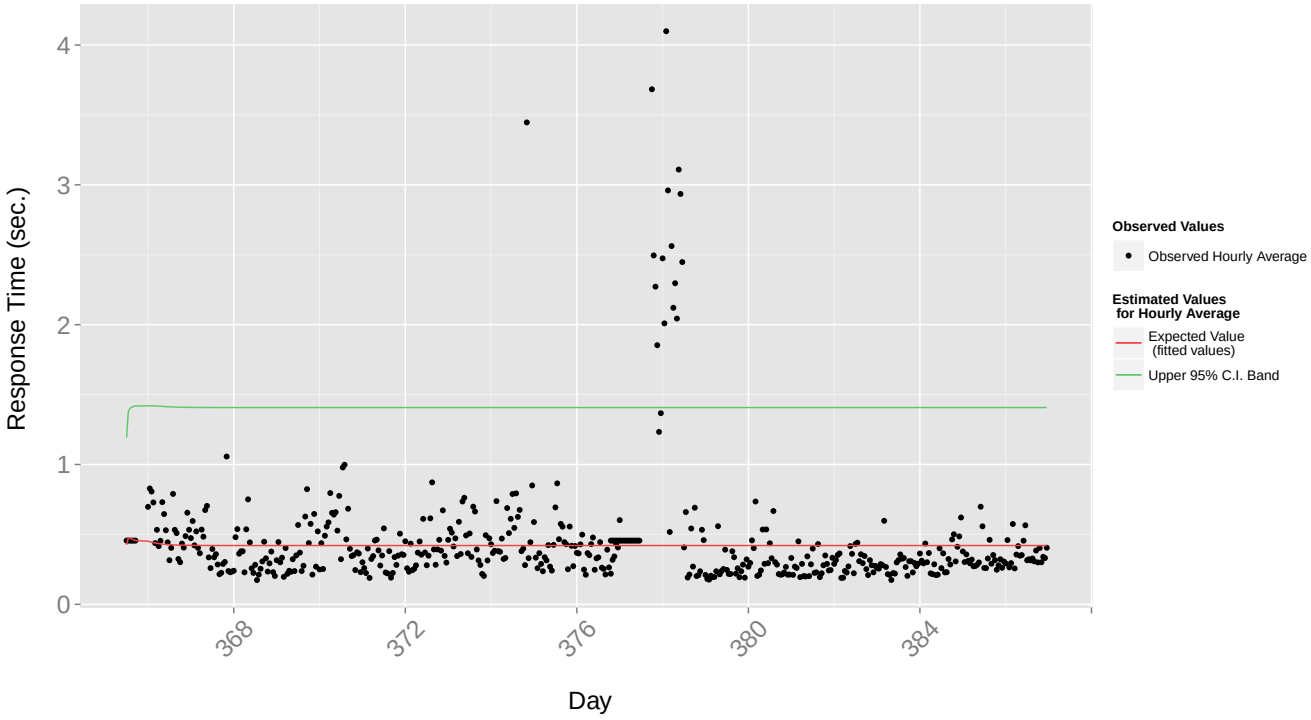


Figure 6.4: A time series plot the observed hourly average response times from Day 366 to Day 386 for the Weblogic c12 Server application. The time series plot contains the estimated hourly average response times. These estimated values are calculated from the forecast function the estimated model parameters from the fitted non-seasonal $ARIMA(2, 2, 0)$ model.

Chapter 7

Conclusion

The theoretical minimum response time of an application serves as a basis for the application's expected response time. We have shown through the Oracle eBusiness Suite and Weblogic c12 Server applications that estimating the daily minimum response time results in an estimated lower limit. This estimated lower limit represents the minimum amount of time that the application should take to complete any transaction on that day. When an application's response time is greater than a certain threshold, there is likely an anomaly in the application that is causing unusual performance issues. We have also shown that estimating the distribution of the theoretical average response time can be used to calculate the value of this threshold.

Since Sandia National Laboratory's Application Services and Analytics departments middleware services provide support to the entire laboratory, it is important that we research and implement analytic capabilities that can improve our understanding of an applications transactional response time. Therefore, it is beneficial to implement the findings of this research to all types of transactional applications that are being used by the Application Services and Analytics departments. Implementing these methods will result in quicker and more accurate anomaly detection that will lead to better problem management.

References

[1] Canova, F., Hansen, B (1995). 'Are seasonal patterns constant over time? a test for seasonal stability'. Journal of Business and Economic Statistics, (13):237-252.

[2] Casella, G., and Berger, R. (2002). Statistical inference (2nd ed.). Australia: Thomson Learning.

[3] Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer-Verlag, London, U.K.

[4] Haan, L. De, and Ferreira, A. (2006). Extreme Value Theory: An Introduction. New York: Springer.

[5] Heffernan, J and Stephenson, A (2012). ismev: An Introduction to Statistical Modeling of Extreme Values. R package version 1.39. URL http://CRAN.R-project.org/package=ismev.

[6] Hyndman, R. et al (2015). forecast: Forecasting functions for time series and linear models. R package version 5.6. URL http://CRAN.R-project.org/package=forecast.

[7] Hyndman, R. and Khandakar, Y. (2008) "Automatic time series forecasting: The forecast package for R", Journal of Statistical Software, 26(3).

[8] Kwiatkowski D, Phillips PC, Schmidt P, Shin Y (1992). 'Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root'. Journal of Econometrics, 54, 159 - 178.

[9] Peiris, M. and Perera, B. (1988), 'On prediction with fractionally differenced ARIMA models", Journal of Time Series Analysis, 9(3), 215-220.

[10] Prado, Raquel, and Mike West (2010). Time Series: Modeling, Computation, and Inference. Boca Raton: CRC.

[11] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[12] West, Mike. Electroencephalogram Recordings. Department of Statistical Science Duke University.

[13] Wickham, Hadley (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29.

Appendix A

Hyndman and Khanadakar ARIMA Model Selection Algorithm

The following is a summary of the Hyndman and Khanadakar ARIMA model selection algorithm that can be found in the Statistical Journal, "Automatic time series forecasting: The forecast package for R". The algorithm is used in the *auto.arima*() function in R [7].

A.0.1 Non-Seasonal: $s = 1$

Saturated Model: $ARIMA(p, q, d)$

Step 1:

Select d using the KPSS unit-root test explained in section 2.2.1. This involves increasing the value of d until the KPSS unit-root test returns an insignificant result. The first value of d that returns an insignificant result will be selected for the model.

Step 2:

Fit the following four models using the function arima() in R programming:

1. $ARIMA(2, d, 2)$
2. $ARIMA(0, d, 0)$
3. $ARIMA(1, d, 0)$
4. $ARIMA(0, d, 1)$

If $d \leq 1$ then these models are fit with $c \neq 0$. Select the model that has the smallest AIC value, where $AIC = 2(p + q + K) - 2ln(L)$,

and where $k = 1$ if $c \neq 0$ and L is the maximized likelihood of the model fitted to the non-seasonal differenced data $D^d y_t$. The selected model will be called the "current" model.

Step 4:

Fit the following variations of the "current" model from Step 3 where:

1. One of the model orders p and q varies by ± 1 from the current model.
2. Both model order p and q vary by ± 1 from the current model.
3. Either include or exclude the parameter c, depending on if it's included in the "current" model.

If the AIC value one of the variations of the "current" model has a lower AIC value then the "current" model, then it becomes the new "current model". After all variations have been fitted, the "current" model is selected as the most appropriate model.

In order to avoid problems with convergence or near unit roots, *Hyndman* and *Khandakar* give a list of constraints that can be found in their Journal of Statistical software [7].

A.0.2 Seasonal: $s \neq 1$

Saturated Model: $ARIMA(p, q, d)(P, DQ, D)[s]$

Step 1:

Select D, where $D = 0$ or 1, using the Canova-Hansen Test. The Canova-Hansen Test checks whether the change seasonal pattern of a realization changes sufficiently over time leads to a unit root. Further discussion of this test can be found in Canoca and Hansen 1995 [1].

Step 2:

Select d by applying the KPSS unit-root test, section 2.2.1, to the seasonal differenced data $D_s^D y_t$. This involves increasing the value of d until the KPSS unit-root test returns an insignificant result. The first value of d that returns an insignificant result will be selected for the model.

Step 3:

Fit the following four models using the function arima() in R programming:

1. $ARIMA(2, d, 2)(1, D, 1)[s]$
2. $ARIMA(0, d, 0)(0, D, 0)[s]$
3. $ARIMA(1, d, 0)(1, D, 0)[s]$
4. $ARIMA(0, d, 1)(0, D, 1)[s]$

If $D + d \leq 1$, then these models are fit with $c \neq 0$. Select the model that have the smallest AIC value, where $AIC = 2(p + q + P + Q + k) - 2ln(L)$, and where $k = 1$ if $c \neq 0$ and L is the maximized likelihood of the model fitted to the differenced data, $D_s^d D^d y_t$. The selected model will be called the "current" model.

Step 4:

Fit the following variations of the "current" model from Step 3 where:

1. One of the model orders p, q, P and Q varies by ± 1 from the current model.
2. Both model order p and q vary by ± 1 from the current model.
3. Both model order P and Q vary by ± 1 from the current model.
4. Either include or exclude the parameter c, depending on if it's included in the "current" model.

If the AIC value one of the variations of the "current" model has a lower AIC value then the "current" model, then it becomes the new "current model". After all variations have been fitted, the "current" model is selected as the most appropriate model.

In order to avoid problems with convergence or near unit roots, $Hyndman$ and $Khandakar$ give a list of constraints that can be found in their Statistical Journal, "Automatic time series forecasting: The forecast package for R" [7].

DISTRIBUTION:

1	MS 0805	W.R. Cook, 09530 (electronic copy)
1	MS 0820	J.P. Abbott, 09339 (electronic copy)
1	MS 0838	G.K. Rogers, 09330 (electronic copy)
1	MS 0933	J.O. Gallegos, 09500 (electronic copy)
5	MS 1465	M.R. Paiz, 09533
1	MS 1465	M.R. Paiz, 09533 (electronic copy)
1	MS 1465	G.N. Conrad, 09533 (electronic copy)
1	MS 0899	Technical Library, 9536 (electronic copy)


Sandia National Laboratories